

# Structural studies on Erythrina cristagalli lectin and botulinum neurotoxin A

Turton, Kathryn

*Award date:*
2003

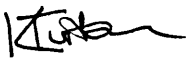
*Awarding institution:*
University of Bath

# STRUCTURAL STUDIES ON *ERYTHRINA CRISTAGALLI* LECTIN AND BOTULINUM NEUROTOXIN A

submitted by Kathryn Turton

for the degree of PhD
of the University of Bath
2003

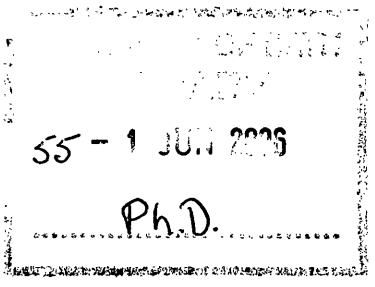








ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346


55 - 1 JUN
Ph.D.

# ABSTRACT

*Erythrina cristagalli* lectin is a galactose-specific legume lectin that specifically recognises dorsal root ganglia (neurons involved in the sensation of pain). Botulinum neurotoxin A, the causative agent of botulism, induces flaccid paralysis by inhibiting the release of acetylcholine at the neuromuscular junction. *Erythrina cristagalli* lectin can be used to retarget the potent activity of fragments of botulinum neurotoxin A to dorsal root ganglia. Conjugates composed of *Erythrina cristagalli* lectin chemically coupled to non-toxic fragments of botulinum neurotoxin A have potential therapeutic utility in the treatment of chronic pain syndromes. It was anticipated that structural analysis of the protein components would shed light on their biological function and possibly allow rational improvement of the design of the conjugate.

The structure of natively-sourced *Erythrina cristagalli* lectin was reported in 2002 in complex with lactose and fucosyllactose (Svensson et al 2002). When this report was published, the structure of *Erythrina cristagalli* lectin was also being analysed as part of this investigation. Data had been collected on the crystal structures of native and recombinant forms of *the* lectin, both unliganded and in complex with lactose. There are some differences between the published structure and the structure of the native lectin determined in this investigation at the amino acid sequence level. The tertiary structure of *Erythrina cristagalli* lectin was confirmed as being homologous to the known structures of other legume lectins, although protomers dimerise in a noncanonical fashion. Lactose is recognised via the galactose ring and forms a network of hydrogen bonds and hydrophobic interactions with combining site residues $Leu^{86}$, $Asp^{89}$, $Tyr^{106}$, $Gly^{107}$, $Phe^{131}$, $Asn^{133}$, $Ala^{218}$ and $Gln^{219}$. Except for the absence of glycosylation from recombinant ECL, there are no differences between the structures of native and recombinant forms of the lectin.Since glycosylation has no role in dimerisation or carbohydrate recognition, native and recombinant forms of *Erythrina cristagalli* lectin can be considered equivalent.

Two non-toxic fragments of botulinum neurotoxin A were studied – one active and one catalytically inactive. The three-dimensional structures of the toxin fragments could not be determined because crystallisation was unsuccessful, so their secondary structures were analysed using circular dichroism spectroscopy. The fragments were highly similar, in terms of their secondary structures, comprising approximately 33% α-helix, 18% β-strand and 31% random coil.

# ACKNOWLEDGEMENTS

Firstly, I would like to thank my two supervisors, Professor Ravi Acharya and Dr John Chaddock, for their help and advice during my postgraduate studies. I am grateful to have had the opportunity to be involved in a project with potential therapeutic benefit and to learn the techniques involved in protein X-ray crystallography. I would also like to offer thanks to the Department of Biology & Biochemistry at the University of Bath and the Health Protection Agency, formerly the Centre for Applied Microbiology & Research (CAMR), at Porton Down for funding my research.

I am grateful to all my colleagues, past and present, at the University of Bath for teaching me the practical aspects of X-ray crystallography. In particular, I would like to thank Dr Matthew Baker, Dr Gayatri Chavali, Dr Evangelia Chrysina, Dr Susan Crennell, Michelle Hares, Dr Daniel Holloway, Dr Shalini Iyer and Dr Jawahar Swaminathan for their help and advice in the laboratory and Dr Ramanathan Natesh for modelling the glycosylation on native ECL. I would also like to thank Dr Laurie Irons for all the supervision, assistance and information he provided regarding the circular dichroism experiments and subsequent analysis.

In addition, I would like to thank the members of the Toxin Therapeutics Division at CAMR for providing pure protein for this investigation and for answering my queries about their expression and purification procedures. I would also like to acknowledge staff at CCLRC Daresbury Laboratory, Daresbury for technical support with the Synchrotron Radiation Source.

Finally, I would like to dedicate this thesis to my husband, whose love and encouragement has kept me motivated over the last three years, and to my parents, without whom none of this would have been possible. I would also like to thank my sister and Grandmas, who have supported me through my studies.

# DEDICATION

For my family,

"Begin at the beginning"..."and go on till you come to the end: then stop."

from "Alice's Adventures in Wonderland"
by Lewis Carroll (first published in 1865)

# ABBREVIATIONS

| | |
|---|---|
| $\sigma$ | Estimated error |
| $\varepsilon$ | Extinction coefficient |
| I | Intensity |
| $\alpha$ | Phase angle |
| $\lambda$ | Wavelength |
| $\theta$ | Angle of diffraction |
| $\rho$ | Electron density |
| °C | Degrees Celsius |
| µl | Microlitre |
| \|F\| | Structure factor amplitude |
| Å | Ångstrom |
| BoNT | Botulinum neurotoxin |
| CCD | Charge-coupled device |
| CCF | Correlation coefficient |
| CD | Circular dichroism |
| CGRP | Calcitonin gene-related peptide |
| cm | centimetre |
| CNS | Central nervous system |
| CNT | Clostridial neurotoxin |
| ConA | Concanavalin A |
| d | Crystal spacing |
| Da | Dalton |
| DRG | Dorsal root ganglia |
| DT | Diphtheria toxin |
| ECL | *Erythrina cristagalli* lectin |
| ECorL | *Erythrina corallodendron* lectin |
| F | Structure factor |
| $F_{calc}$ or $F_C$ | Calculated structure factor amplitudes |
| $F_{obs}$ or $F_O$ | Observed structure factor amplitudes |
| Fuc | Fucose |
| Gal | Galactose |

| | |
|---|---|
| GalNAc | N-acetyl-D-galactosamine |
| GlcNAc | N-acetyl-D-glucosamine |
| GSIV | *Griffonia simplicifolia* lectin 4 |
| h,k,l, | Miller indices |
| $H_C$ | Binding domain |
| HEPES | N-(2-Hydroxyethyl)piperazine-N'-(2-ethanesulphonic acid) |
| $H_N$ | Translocation domain |
| K | Kelvin |
| kDa | KiloDalton |
| kg | Kilogram |
| kV | Kilovolts |
| Lac | α-lactose |
| LacNAc | N-acetyllactosamine |
| LC | Light chain |
| LenL | Lentil lectin |
| $LH_N/A$ | Non-toxic fragment of BoNT/A |
| $LH_N/A(H^{227}Y)$ | Catalytically inactive mutant of $LH_N/A$ |
| M | Molar |
| MAD | Multiwavelength anomalous dispersion |
| Man | Mannose |
| mg | Milligram |
| MIR | Multiple isomorphous replacement |
| ml | Millilitre |
| mm | Millimetre |
| mM | Millimolar |
| MPD | 2-methyl-2,4-pentanediol |
| NAP | Neurotoxin associated protein |
| *n*ECL | Native ECL |
| ng | Nanogram |
| nm | Nanometre |
| NSF | N-ethylmaleimide sensitive factor |
| PAGE | Polyacrylamide gel electrophoresis |
| PDB | Protein Data Bank |
| PEG | Polyethylene glycol |
| PHA-L | Phytohaemagglutinin |
| PNA | Peanut agglutinin |

| | |
|---|---|
| PSL | Pea lectin |
| $R_{cryst}$ | Crystallographic R factor |
| *rec*ECL | Recombinant ECL |
| RF | R factor |
| $R_{free}$ | Free R factor |
| RMSD or rmsd | Root mean square deviation |
| $R_{symm}$ | R factor comparing symmetry-related reflections |
| s | Second |
| SBA | Soybean agglutinin |
| SDS | Sodium dodecylsulphate |
| SIR | Single isomorphous replacement |
| SNAP-25 | Synaptosome-associated protein of 25kDa |
| SNARE | Soluble NSF-attachment protein receptor |
| SRS | Synchrotron Radiation Source |
| TeNT | Tetanus neurotoxin |
| Tris | Tromethamine |
| VAMP | Vesicle-associated membrane protein |
| Vm | Matthews coefficient |
| WBA-1 | Winged bean acidic lectin |
| WGA | Wheat germ agglutinin |
| x,y,z | Atomic coordinates |
| Xyl | Xylose |

# CONTENTS

# CHAPTER ONE
# INTRODUCTION

Non-toxic fragments of botulinum toxin A conjugated to *Erythrina cristagalli* lectin are of potential therapeutic use in the treatment of chronic pain syndromes (Duggan et al. 2002). In addition to information on the biological activity of these conjugates *in vitro* and *in vivo*, the structures of the components are of interest as they can provide insight into the mechanism of action and might allow the conjugates to be specifically redesigned to improve their function.

## 1.1 SENSATION AND TRANSMISSION OF PAINFUL STIMULI

Nociception, the sensation of pain, is triggered by the stimulation of peripheral nerves and is normally associated with tissue damage. Peripheral sensory neurons transmit painful stimuli from the site of injury to the central nervous system (CNS). These primary sensory neurons (known as nociceptors) can be stimulated by noxious stimuli (for example, intense heat or pressure) but not by innocuous stimuli, such as gentle warming or light touch (Julius and Basbaum 2001).

There are three main groups of sensory neurons, Aβ-, Aδ- and C-fibres. Aβ-fibres are myelinated, rapidly conducting neurons with large cell bodies that detect innocuous stimuli. Aδ-fibres are thinly myelinated, rapidly conducting neurons that mediate the sensation of rapid, acute, sharp pain. C-fibres are unmyelinated, slow-conducting fibres that signal the delayed, more diffuse, dull painful stimuli. In all nociceptors, glutamate is the predominant excitatory neurotransmitter (Julius and Basbaum 2001), although the peptide neurotransmitters substance P and calcitonin gene-related peptide (CGRP) are also involved.

Dorsal root ganglia (DRG) are nociceptive afferents that transmit peripheral noxious stimuli to spinal cord neurons at the dorsal horn. Noxious signals, transmitted from the DRG by the release of glutamate, substance P and CGRP, are integrated in the spinal cord and transferred on to the thalamus before reaching the somatosensory cortex where they are recognised as painful stimuli (Riedel and Neeck 2001).

### 1.1.1 Chronic Pain

Pain is an unpleasant sensation that is normally associated with tissue damage. Chronic pain is caused by the continual stimulation of nociceptors in areas of tissue damage but persists long after the injury has healed (Ashburn and Staats 1999).

Common syndromes of chronic pain include myofascial pain, phantom limb pain and chronic low back pain. Chronic pain is thought to affect up to 45% of the population and is one of the most common reasons why people seek medical attention (Elliott et al. 1999). Effective treatment involves a number of therapies because the patient's general health is affected: people with chronic pain commonly also suffer depression, sleep disturbance and fatigue (Ashburn and Staats 1999). In terms of financial burden, chronic pain is an expensive public health problem because of the number of working days lost each year and the cost of long-term pharmaceutical treatments.

Several pharmaceuticals are available for the treatment of chronic pain although some induce unpleasant side effects and none are entirely effective. Non-steroidal anti-inflammatory drugs, for example, can cause gastrointestinal irritation and ulceration (Allison et al. 1992) as well as renal dysfunction (Clive and Stoff 1984). Opioid analgesics, such as morphine, can effectively control chronic pain but their use is associated with constipation, sedation, tolerance and addiction (Ashburn and Staats 1999).

## 1.2 *ERYTHRINA CRISTAGALLI* LECTIN

*Erythrina cristagalli* lectin (ECL) is a galactose-specific lectin isolated from seeds of the leguminous coral tree *E.cristagalli*. ECL is biologically active as a heterodimer of approximately 56.8kDa molecular weight, comprising two glycosylated subunits of 28kDa and 26kDa (Iglesias et al. 1982). Its biological function in the legume is currently unknown, although, *in vitro*, ECL has been demonstrated to agglutinate rabbit and human erythrocytes (Iglesias et al. 1982). ECL is also mitogenic for human peripheral blood T lymphocytes and can potentially be used to differentiate between cell types according to the oligosaccharides expressed on the cell surfaces. The crystal structure of ECL was first reported in 2002 (Svensson et al 2002). This published structure (herein referred to as *Sven*ECL) differs from the crystal structures of native and recombinant ECL determined in the current investigation, as described and discussed in Chapter 3.

## 1.3 BOTULINUM NEUROTOXIN A

Botulinum neurotoxin A (BoNT/A) is one member of a family of toxins produced and secreted by bacteria of the genus *Clostridium*. The botulinum neurotoxins are the most poisonous proteins known to Man. They cause the disease botulism, which is characterised by a descending flaccid paralysis resulting from the blockade of acetylcholine release at the neuromuscular junction.

BoNT/A is a di-chain molecule in which the light and heavy chains are linked by a disulphide bond. The toxin contains three functional domains, each of which is involved in the mechanism of action. The binding domain enables the toxin to recognise cholinergic neurons and to be internalised into an endosome. The translocation domain forms an ion-conducting channel spanning the endosomal membrane through which the catalytic domain translocates to the cytosol. The catalytic domain has zinc-dependent endopeptidase activity and acts upon the SNARE protein SNAP-25, thus preventing the release of neurotransmitter molecules into the synapse.

BoNT/A has been licensed as a biopharmaceutical since 1989 and is currently available in two preparations: Botox (Allergan Inc, USA) and Dysport (Speywood Pharmaceuticals Ltd, UK). Due to its action in cholinergic neurons, BoNT/A can be used to relieve a variety of muscle spasm disorders as well as complaints of the autonomic nervous system. In conjunction with relief of spasticity, the use of BoNT/A is associated with antinociception, which has led to its proposed utility as a therapeutic agent for the treatment of chronic pain syndromes, such as migraine and tension headaches (Aoki 2003; Wheeler 1998). In particular, BoNT/A is noted for its longevity of action (the induced paralysis lasts for up to 6 months) although there are side effects associated with its use and the possibility of developing immunity to the toxin. More recently, fragments of BoNT/A and other clostridial neurotoxins have been utilised as delivery vehicles for specifically targeting ligands to cholinergic neurons. In addition, the cell-binding domains of these toxins has been replaced with cell-binding moieties of alternative specificity, enabling the potent activity of the toxin to be selectively retargeted to non-native targets and increasing the number of potential therapeutic applications of these proteins.

## 1.4 RETARGETING THE ACTIVITY OF BOTULINUM NEUROTOXIN A

A non-toxic derivative of BoNT/A that lacks the cell-binding domain has been prepared (Chaddock et al. 2002). This fragment, known as $LH_N/A$, is unable to bind neurons but retains the full endopeptidase activity of the native toxin. Chemically coupling $LH_N/A$ to non-native cell-binding moieties provides a means of specifically retargeting the catalytic activity of BoNT/A to cells other than cholinergic neurons. For example, when applied to a variety of cell lines (including those normally resistant to the effects of BoNT/A) a conjugate of $LH_N/A$ and wheat germ agglutinin was found to inhibit secretory responses (Chaddock et al. 2000b). More recently, $LH_N/A$ has been coupled to ECL, forming the $LH_N/A$-ECL conjugate, which specifically targets DRG (Duggan et al. 2002). Unlike other neurons, DRG express galactose-containing

oligosaccharides on their surfaces (Streit et al. 1985; Streit et al. 1986), which can be recognised by ECL (Duggan et al. 2002).

The action of $LH_N$/A-ECL on DRG inhibits the release of glutamate, substance P and CGRP, thereby preventing the transmission of painful stimuli to the spinal cord (Duggan et al. 2002). The longevity of action of the conjugate is equivalent to that of the native toxin and its action does not affect non-noxious activity. Application of the conjugate to rats has provided evidence that it could become a useful therapeutic agent in the treatment of chronic pain (Duggan et al. 2002). Moreover, the side effects associated with conventional analgesics are not predicted for $LH_N$/A-ECL.

## 1.5 AIMS OF THE INVESTIGATION

This investigation has formed part of collaboration between the University of Bath and the Health Protection Agency, where the $LH_N$/A-ECL conjugate has been prepared and tested. The ultimate aims of the project were to determine and analyse the three-dimensional structures of $LH_N$/A and ECL. It was anticipated that structural knowledge of the protein components of the $LH_N$/A-ECL conjugate would provide insight into its biological function and possibly allow rational improvement of the design of the conjugate. The method of choice was X-ray crystallography, which is routinely used to accurately determine the structures of biological macromolecules to high resolution. Protein structures have been determined crystallographically since the 1950s and there are now more than 22,400* protein structures deposited in the Protein Data Bank, PDB (Berman et al. 2000; www.rcsb.org/pdb/).

### 1.5.1 Structural Studies on Fragments of Botulinum Neurotoxin A

The effects of $LH_N$/A on dorsal root ganglia were compared with those of a catalytically inactive fragment, known as $LH_N$/A($H^{227}Y$) in which $His^{227}$, a key residue in the catalytic mechanism of the toxin, was mutated to tyrosine (Duggan et al. 2002). Mutating $His^{227}$ to Tyr abolishes the catalytic activity of BoNT/A (Zhou et al. 1995), making $LH_N$/A($H^{227}Y$) a useful control agent for assessing the non-catalytic effects of the conjugate on target cells. Therefore, the initial focus of the investigation was to crystallise both $LH_N$/A and $LH_N$/A($H^{227}Y$) and then to determine and compare their crystal structures.

Since the crystal structure of BoNT/A is known (Lacy et al. 1998), atomic coordinates for the catalytic and translocation domains were to be used as a search model in molecular replacement to determine initial phases for the toxin fragments.

* 22,448 protein structures deposited by 9 September 2003

There is only one amino acid different between the primary structures of the two fragments, but as this mutation abolishes the catalytic activity of $LH_N/A(H^{227}Y)$, any structural rearrangements in the active site were to be analysed. The final structural models for each fragments were then to be compared with the known crystal structure of the BoNT/A holotoxin (PDB code 3BTA) to assess any structural changes brought about by removal of the binding domain.

The BoNT/A translocation domain undergoes structural rearrangement at low pH, forming a membrane-spanning channel that facilitates release of its light chain into the neuron cytosol. In order for the catalytic light chain to be activated, the disulphide bond linking it to the translocation domain must be reduced. Currently, it is not known at what point during the mechanism of action this takes place, but it is known that the structure of the light chain after translocation is different to its structure when joined to the toxin heavy chain (Li and Singh 2000b). Therefore, the nature of any structural changes resulting from reduction of the disulphide bond were to be investigated by co-crystallising $LH_N/A$ with dithiothreitol, or by adding dithiothreitol to pre-formed crystals prior to data collection. It was hoped that these studies would provide insight into the location of the translocation belt following reduction of the interchain disulphide bond.

Furthermore, co-crystallisation of $LH_N/A$ and $LH_N/A(H^{227}Y)$ with substrates or substrate analogues was to be attempted in order to fully understand the effects of mutating $His^{227}$ to Tyr on the catalytic site. Of particular interest, were the mode of zinc coordination in $LH_N/A(H^{227}Y)$ compared to $LH_N/A$ and BoNT/A and chemical details of the substrate-enzyme interface. Potential inhibitors of the metallopeptidase activity of BoNT/B (Adler et al. 1998; Eswaramoorthy et al. 2002) and substrate analogues for BoNT/A, based on truncated versions of SNAP-25, have recently been described (Hallis et al. 1996; Schmidt et al. 1998). Co-crystallisation of the toxin fragments with these compounds might shed light on the catalytic mechanism of BoNT/A.

### 1.5.2 Structural Studies on *Erythrina cristagalli* Lectin

The ECL-$LH_N/A$ conjugate is prepared using a recombinantly over-expressed form of ECL. It is, therefore, necessary to ensure that recombinant ECL (*rec*ECL) is identical in structure and activity to native ECL (*n*ECL) so that it may be expected to behave in the same way in biological systems. The main structural difference between recombinant and native forms of ECL is the absence of glycosylation from the recombinant form, since the expression host (*Escherichia coli*) does not contain the cellular machinery required to glycosylate proteins in the same way as eukaryotic cells. Therefore, the aim was to determine and compare the crystal structures of both native

and recombinant forms of ECL to demonstrate that the structure of the recombinant form does not differ significantly from that of native ECL. The crystal structure of the closely related *Erythrina corallodendron* lectin was available as a search model for molecular replacement.

The crystal structure of *n*ECL was also to be compared with the known structures of other legume lectins to analyse the mode of dimerisation and assess the potential role of glycosylation in the observed mode of quarternary association. Additionally, the quarternary association of *rec*ECL protomers was of interest. *Erythrina corallodendron* lectin, which is closely related to ECL, has an unusual dimer structure that is postulated to be caused by the presence of a bound oligosaccharide. Therefore, it was of interest to determine whether or not ECL adopts the canonical mode of dimerisation and, if not, whether or not the variation could be attributed to the presence of carbohydrate by comparing the quarternary association of native and recombinant (i.e. glycosylated and non-glycosylated) forms.

Half-way through the current investigation, the crystal structures of natively-sourced ECL in complex with lactose and fucosyllactose (PDB codes 1GZ9 nd 1GZC) were reported (Svensson et al 2002). These data showed that *Sven*ECL protomers adopt the jelly-roll fold and dimerise back-to-back in a non-canonical fashion. N-linked glycosylation was observed at two positions in *Sven*ECL, Asn17 and Asn113, and the side-chains of Asp89, Gly107 and Asn133 were shown to create a strong hydrogen-bonding network for carbohydrate binding.

In addition, co-crystallisation of both *n*ECL and *rec*ECL with a variety of galactose-containing saccharides was to be attempted in order to glean information on the architecture of the combining site and the mode of carbohydrate binding. Analysis of the combining site was of interest because the activity of ECL in retargeting $LH_N/A$ to DRG is based upon its carbohydrate specificity. Although DRG are the only neurons expressing galactose-containing carbohydrates on their surfaces, the specificity of the interaction might be enhanced by site-directed mutagenesis if we understand the architecture of the combining site and have knowledge of the residues involved in carbohydrate binding. Furthermore, despite high sequence similarity, slight differences have been noted in the carbohydrate-binding properties of ECL compared with other *Erythrina* lectins. The basis of these differences is also of interest, as it will provide insight into the variation in carbohydrate specificity observed among lectins with similar tertiary and quarternary structures.

# CHAPTER TWO
# PROTEIN CRYSTALLOGRAPHY

Crystallography, the science of crystal structures, properties and forms, can be utilised to determine the three-dimensional structures of biological macromolecules. Determination of protein structures by this method involves collecting and analysing the diffraction patterns produced when protein crystals scatter X-rays. Due to their short wavelengths, X-rays allow resolution down to the atomic level. A beam of X-rays incident on a protein crystal will be scattered and the resultant diffraction pattern can be recorded. No physical X-ray lens exists, so recombination of the scattered X-rays must be simulated using information on the intensity and phase of each diffracted X-ray. However, only the intensities of the X-rays can be recorded in the diffraction pattern, creating a "phase problem", which must be overcome in order to elucidate structural information.

The crystallographic determination of protein structures involves a series of experimental steps: The protein must be isolated, purified and crystallised before X-ray diffraction data can be collected. The recorded data must then be processed and initial phases determined in order to calculate electron density maps of the protein structure. A model is then fitted to the electron density and repeated cycles of refinement and model building are performed until the model cannot be further improved.

## 2.1 OBTAINING PURE PROTEIN

Before crystallisation trials can commence, the protein of interest must be isolated and purified. Proteins can either be obtained directly from a wild type source (referred to as "native") or recombinantly over-expressed in a host organism. Milligram quantities of protein are required for crystallisation and the protein should be soluble and in the authentic conformation as well as being of the highest purity.

### 2.1.1 Hosts for the Recombinant Expression of Proteins

It can be expensive and impractical to isolate protein from wild type sources, particularly if the protein of interest is mammalian, so the gene coding for the protein is often introduced into a host cell system for over-expression. Several systems can be used for the preparation of recombinant proteins – bacteria, yeasts, insect cells and mammalian cells. Each expression system has its own relative merits and associated disadvantages and no one system is suitable for all practical applications.

*Escherichia coli*, the most common initial choice for protein over-expression, has several advantages over other hosts. The bacterium has been well characterised in terms of its genetics, molecular biology and physiology and is easy to culture with a high growth rate. It is easy to extract protein from *E.coli*, but the cells have limited post-translational machinery and the fine structure of genes is different to that in higher eukaryotes. Many proteins form insoluble aggregates (inclusion bodies) when over-expressed in *E.coli* because they are incorrectly folded (Georgiou and Valax 1996) but this can be avoided by slowing the rate of protein synthesis; expressing the protein of interest as a fusion with a highly soluble protein; or by co-expressing molecular chaperones. If the protein is toxic to *E.coli* and expression as a fusion or targeting to the periplasm does not reduce the toxicity, another expression host should be considered.

Yeasts, such as *Pichia pastoris* and *Saccharomyces cerevisiae*, are as cheap and easy to use as bacterial expression systems with the added advantage of having post-translational processing machinery similar to mammalian cells (Sudbery 1996). The protein of interest can be directed into the yeast secretory pathway, where protein folding, disulphide bond formation and glycosylation occur, although the nature of the glycosylation can differ from that of the native protein. Insect cells (for example, from *Drosophila melanogaster* and *Spodoptera frugiperda*) can be used as expression systems in conjunction with baculovirus or plasmid vectors (McCarroll and King 1997). Insect cells are cheap to maintain and can perform the same types of post-translational modification as mammalian cells. Mammalian cells (e.g. Chinese hamster ovary cells) provide the most native environment for the over-expression of mammalian proteins and can produce recombinant proteins that have near-native glycosylation profiles and post-translational modifications.

### 2.1.2 Methods of Protein Purification

The protein of interest must be purified, whether it has been isolated from a wild-type source or recombinantly over-expressed in a host system, to remove other protein species and biomacromolecules. Purity is important for crystallographic studies and great care must be taken to prepare material that is at least 99% pure protein for crystallisation trials. Affinity purification typically provides a protein sample that is 80-90% pure, with further purification steps normally including ion-exchange and size-exclusion chromatography (Lesley 2001). Incorporation of foreign molecules is likely to introduce lattice defects, inhibiting crystal growth. The solution should also be monodisperse (i.e. the protein should exist as a single, oligomeric species) and

conformationally pure because differences in the charge or glycosylation of protein molecules, for example, can hinder crystallisation.

Ion-exchange chromatography separates proteins according to differences in charge of the protein species. The impure protein sample is loaded onto a column that is oppositely charged to the protein of interest. Conditions (e.g. salt concentration) are then altered stepwise to elute the bound protein species. Affinity chromatography separates proteins using reversible affinity for a specific ligand that is attached to the column. The rationale is that only the protein of interest will bind to the ligand and once all unbound protein has been removed, the protein can be released from the column by applying another ligand for which the protein has higher affinity or by altering the pH, ionic strength or polarity of the column. Gel filtration chromatography is used to separate proteins on the basis of their molecular size. This is typically the last step in the purification procedure.

It is vital to ensure that the protein is folded and active after purification and to use fresh stocks for crystallisation so that the protein molecules have not degraded.

## 2.2 CRYSTALLISATION

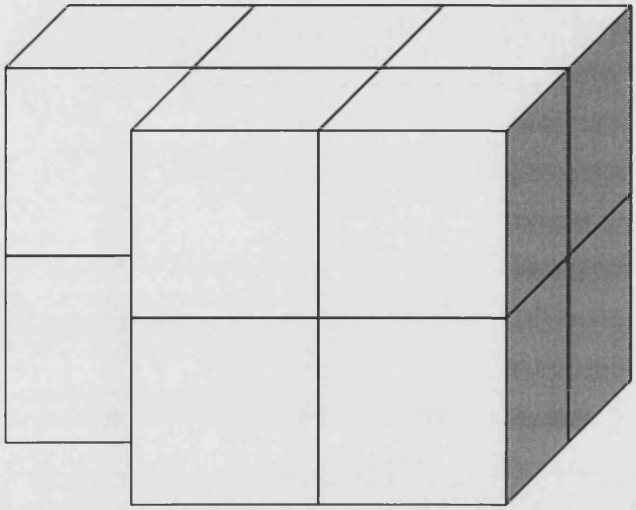
The first step in protein crystallography is crystallisation of the protein of interest. Good-quality, large, single crystals are sought for diffraction experiments because the accuracy of the final structural model depends upon recording good-quality diffraction data. Crystals of proteins differ from those of small molecules in their physical properties. In addition, the crystallisation of proteins is different to that of small molecules because proteins have different physicochemical properties and a larger number of parameters are involved (Drenth 1999). Protein crystals tend to be smaller (less than $10mm^3$) and more fragile with high solvent contents and large unit cells (Ducruix and Giege 1999). Protein crystals are also sensitive to external conditions and are weakly birefringent under polarised light.

### 2.2.1 Crystals and Symmetry

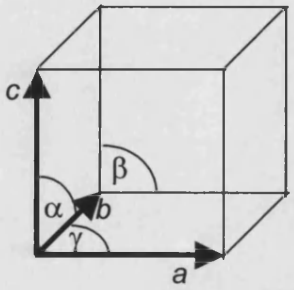
A crystal is a three-dimensional structure in which molecules are repeated at regular intervals in each dimension. During crystallisation, molecules aggregate together in symmetric patterns. This is because the crystallisation conditions favour a small number of intermolecular contacts, which are repeated to form a symmetric pattern. The "unit cell" is the smallest repeating unit that can make up the crystal (by translation operations only) and the information it contains is the structure of all the component molecules. "Crystallographic symmetry" describes the arrangement of molecules within the unit cell. The "asymmetric unit" is defined as the smallest unit that

can be rotated and translated by the crystallographic symmetry to generate one unit cell. The molecules within one asymmetric unit are related by "non-crystallographic symmetry" because they do not have identical environments (Drenth 1999).

**Figure 1 - The Arrangement of Unit Cells in the Crystal Lattice**



Cartoon Representation of the Crystal Lattice




The Unit Cell

The crystal lattice describes the regular arrangement of component molecules. The unit cell, the smallest repeatable building block of the crystal lattice, is defined by dimensions a,b,c and angles α,β,γ. The unit cell contains all the information required to describe the structure of the component molecules, although they do not always exist in the centre of the unit cell. Where there is more than one molecule per unit cell, the asymmetric unit is defined as the smallest unit that can generate one unit cell.

The crystal lattice, defined by three axes (*a*,*b*,*c*) and three angles (α,β,γ) (Figure 1), describes the regular spacing of the component molecules (Ducruix and Giege 1999). The symmetry of the intermolecular interactions defines the symmetry of the crystal lattice and the symmetric constraints imposed on the crystal lattice allow all crystals to be classified into seven crystal systems (Table 1) (Giacovazzo 1994). The symmetry operators that define the arrangement of molecules in crystals include rotations, translations, mirror planes and inversion centres, and different combinations of these operators give rise to 230 space groups. However, because of the enantiomeric nature of amino acids, mirror planes, glide planes and inversion centres cannot be applied to protein crystals, so their symmetries are restricted to only 65 space groups. Protein crystals can have 2-, 3-, 4- or 6-fold axes of symmetry (along with the corresponding screw axes) and can include centring symmetry (McRee 1993).

A space lattice is an arrangement of points where each point occurs in exactly the same environment and orientation as every other point (Blundell and Johnson 1976). If the points are located at each corner of the unit cell, the space lattice is described as primitive (P), but if a point is also placed at the centre of each face, the lattice is face-centred (F). If a point is placed in the centre of the unit cell and at each corner, the lattice is known as body-centred (I), whilst the prefix 'C' identifies monoclinic and orthorhombic lattices in which a point is located at each corner of the unit cell and in the centre of the C face. The fourteen lattices that may be generated from the crystal classes using these criteria are known as Bravais lattices.

**Table 1 – The Seven Crystal Systems and 65 Protein Space Groups**

| Crystal System | Symmetry Constraints | Protein Space Groups |
|---|---|---|
| Triclinic | None | P1 |
| Monoclinic | $\alpha = \gamma = 90°$ (b is unique axis)<br>$\alpha = \beta = 90°$ (c is unique axis) | P2; $P2_1$; C2 |
| Orthorhombic | $\alpha = \beta = \gamma = 90°$ | P222; $P222_1$; $P2_12_12$; $P2_12_12_1$; C222; $C222_1$; F222; I222; $I2_12_12_1$ |
| Tetragonal | $a = b$; $\alpha = \beta = \gamma = 90°$ | P4; $P4_1$; $P4_3$; P42; I4; $I4_1$;<br>P422; $P42_12$; $P4_122$; $P4_322$; $P4_12_12$; $P4_32_12$; $P4_222$; $P4_22_12$; I422; $I4_122$ |
| Trigonal (Rhombohedral) | $a = b$; $\alpha = \beta = 90°$; $\gamma = 120°$<br>($a = b = c$; $\alpha = \beta = \gamma$) | P3; $P3_1$; $P3_2$; R3;<br>P312; P321; $P3_121$; $P3_221$; $P3_112$; $P3_212$; R32 |
| Hexagonal | $a = b$; $\alpha = \beta = 90°$; $\gamma = 120°$ | P6; $P6_1$; $P6_2$; $P6_3$; $P6_4$; $P6_5$;<br>P622; $P6_122$; $P6_522$; $P6_222$; $P6_422$; $P6_322$ |
| Cubic | $a = b = c$; $\alpha = \beta = \gamma = 90°$ | P23; F23; I23; $P2_13$; $I2_13$<br>P432; $P4_132$; $P4_332$; $P4_232$; F432; $F4_132$; I432; $I4_132$ |

All crystals can be classified into one of the seven crystal systems. These crystal systems are defined by the symmetry constraints they impose on the crystal lattice. Different combinations of symmetry operators can be applied to the molecules within a crystal, giving rise to 230 space groups, but the stereochemistry of amino acids limits protein crystals to only 65 of these space groups.

### 2.2.2 Crystallogenesis

Crystallisation is a multi-step process in which nucleation is followed by crystal growth. Nucleation and growth depend upon supersaturation of the protein solution (Weber 1997). Supersaturation is a function of protein concentration and factors such as pH, temperature and ionic strength that affect protein solubility. Protein solutions

can become supersaturated by the alteration of some internal parameters and/or by the addition of precipitation agents (Zanotti 1994). In order to grow large, single crystals for diffraction experiments, supersaturation should be achieved slowly. The methods available for protein crystallisation bring the protein solution into a supersaturated state that favours nucleation and growth.

Crystallogenesis proceeds by the spontaneous association of aggregates in a process known as nucleation (Weber 1997). Formation of nuclei reduces the level of supersaturation of the protein solution, which in turn lowers the probability that further nuclei will form. The degree of supersaturation required for crystal growth is lower than that needed for nucleation and too many nuclei would form if the level of supersaturation remained high (Drenth 1999). Once the nuclei have reached a critical size, crystal growth continues by the addition of protein molecules to the crystalline lattice (Weber 1997). Crystal growth does not continue indefinitely and cessation of growth has several causes, including depletion of protein molecules from the crystallising solution; poisoning of the crystal faces; ageing of the constituent protein molecules; and the presence of growth defects in the crystal lattice.

### 2.2.3 Crystallisation Methods

The search for crystallisation conditions is a somewhat trial-and-error process. Parameters commonly varied include the buffer, pH, precipitant, protein and precipitant concentrations and incubation temperature (Drenth 1999). The presence of additives in low concentrations can also aid crystallisation. With a large number of precipitants, buffers and additives to sample, "sparse matrix" screens are often used to identify initial crystallisation conditions from a wide range of solutions. Subsequent optimisation of conditions that produce initial crystals may then lead to the growth of crystals suitable for diffraction experiments.

Vapour diffusion methods manipulate the concentration of the protein solution remotely, by diffusion through air (Weber 1997). A droplet containing both protein and precipitant, commonly in a 1:1 ratio, is sealed over a reservoir containing the precipitant or crystallisation buffer (also called the "mother liquor"). Hanging drops are dispensed onto a coverslip that is then inverted and sealed over the precipitant. Sitting drops are placed in microbridges that elevate the drop above the level of precipitant in the well. In either case, equilibration leads to the movement of water molecules between the drop and the reservoir, reducing the volume of the drop. The concentration of the protein increases, supersaturating the drop and leading to nucleation and growth.

Crystallisation by dialysis involves containing the protein solution within a "dialysis button" such that it is separated from the mother liquor by a semi-permeable membrane (Ducruix and Giege 1999). Small molecules are free to cross the membrane but protein molecules are not, so the pH and ionic strength of the protein solution can be manipulated directly. Dialysis is a versatile technique because the mother liquor can easily be changed several times, or the concentration of a precipitant increased in incremental stages.

Batch methods involve mixing protein with precipitant at a concentration that will instantly reach supersaturation (Ducruix and Giege 1999). The mixture is sealed and incubated to allow crystal growth. This process has been automated and can handle very small drops, known as "microbatch" crystallisation. In microbatch, the protein and precipitant solutions are mixed and incubated under paraffin oil to prevent evaporation of the precipitant (Chayen 1997). However, the microbatch technique is not suitable for use with organic solvents that are soluble in paraffin oil.

## 2.3 X-RAY DIFFRACTION

A beam of X-rays incident on a protein crystal will be scattered (diffracted) due to interactions with electrons in the crystal lattice. An electron excited by an X-ray will oscillate at the same frequency as the incident wave. This oscillation deflects the X-ray, producing a diffracted beam with the same wavelength as the incident radiation, but differing in phase by 180° (Blundell and Johnson 1976). A crystal contains a large number of electrons, each of which will deflect the incident radiation, and the scattered X-rays will be prone to interference.

The distance between evenly spaced parallel planes of atoms in the crystal is known as the crystal spacing (d). X-rays incident on a plane of atoms at an angle $\theta$ will be diffracted through $2\theta$. However, X-rays diffracted by successive planes of atoms are subject to interference. If the difference in path length between two waves diffracted by successive planes is equal to a whole number of wavelengths, they will interfere constructively and a diffraction spot will be recorded (Figure 2). This is described by Bragg's law:

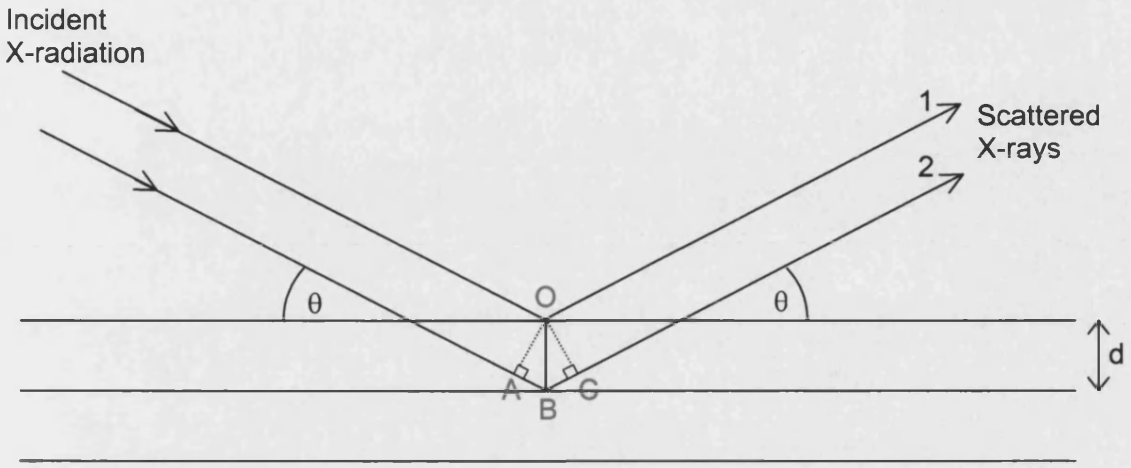
$$n\lambda = 2d \sin\theta$$

where: $\lambda$ is the wavelength of the incident radiation

d is the crystal spacing

$\theta$ is the angle of diffraction

**Figure 2 - Derivation of Bragg's Law**




This diagram represents a crystal lattice in which planes of atoms are separated by a regular distance, d. A beam of X-rays incident on a plane of atoms at an angle θ will be reflected and emerge at an angle of the same value. The difference in path length between X-rays 1 and 2 as they travel through the crystal is equal to AB+BC. If this path difference is equal to a whole number of wavelengths, Bragg's law is satisfied and the emergent X-rays will interfere constructively, giving rise to a reflection in the diffraction pattern.

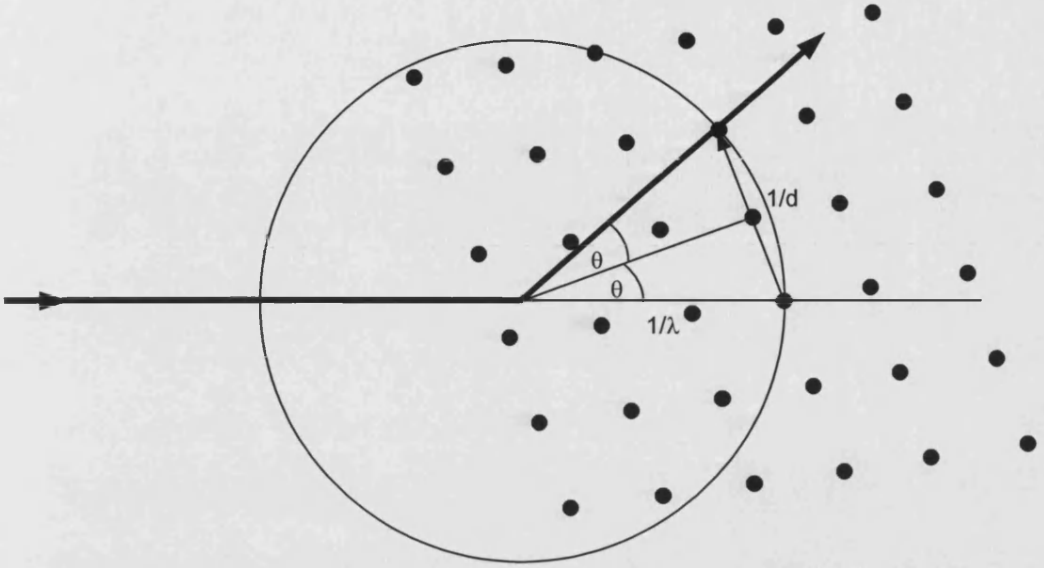
In crystallography, the term "resolution" refers to the crystal spacing. It is related to the scattering angle, θ, as determined by rearranging Bragg's law:

$$d = \frac{0.5}{\sin(\theta)/\lambda}$$

### 2.3.1 The Ewald Construction

The Ewald construction (Figure 3) is a three-dimensional representation of Bragg's law based on the reciprocal lattice which has dimensions inversely proportional to those of the crystal lattice. The incident X-radiation is depicted as a sphere of radius $1/\lambda$ positioned centrally on the beam (Dauter 1999). The reciprocal lattice is oriented with its origin at the point where the X-ray beam passes through the sphere. When a reciprocal lattice point lies on the surface of the Ewald sphere, Bragg's law is satisfied, giving rise to a diffracted beam. This is because each reciprocal lattice point lies at the end of a vector of length 1/d, perpendicular to the corresponding family of crystal planes. Thus, if a reciprocal lattice point lies on the Ewald sphere, $1/2d = (1/\lambda)\sin\theta$, which takes the form of Bragg's law when rearranged.

**Figure 3 - The Ewald Construction**

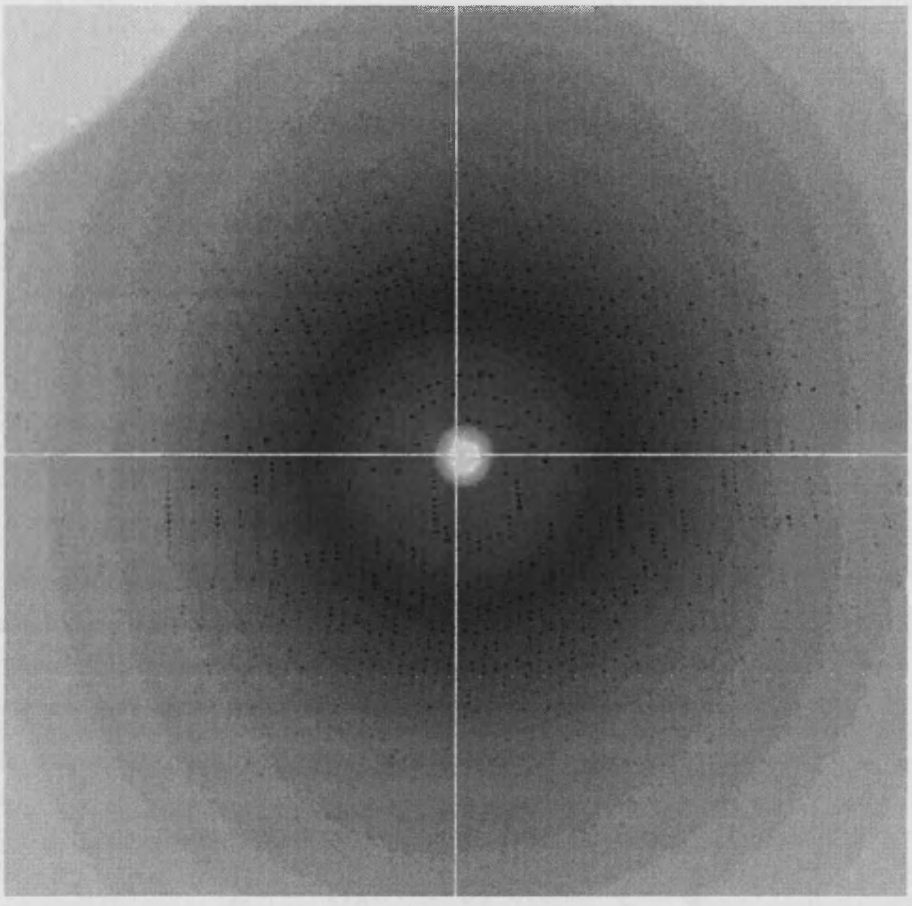



The Ewald construction represents Bragg's law in three dimensions. The incident radiation is shown as a sphere of radius 1/λ (the Ewald sphere) that is positioned centrally on the beam. The reciprocal lattice is oriented with its origin at the point where the X-ray beam passes through the sphere. When a reciprocal lattice point lies on the surface of the Ewald sphere, Bragg's law is satisfied and a reflection will be observed.

Each reciprocal lattice point (i.e. each spot in the diffraction pattern) arises from one diffracted X-ray beam (Drenth 1999).

The diffraction pattern (Figure 4) contains information on the structure of the crystal and its component molecules, which in the present case are protein. The spacing between reflections is inversely related to the crystal spacing because the dimensions of the reciprocal lattice are inversely proportional to those of the crystal lattice. Thus, the smaller the crystal spacing, the greater the distance between lines of reflections in the diffraction pattern. The circular arrangements of spots in the diffraction pattern are known as "lunes". The reflections within each lune arise from diffraction by one plane of the reciprocal lattice. Each reflection corresponds to the molecular transform at a given reciprocal lattice point and its intensity is related to the strength of the molecular transform at that point (Blundell and Johnson 1976).

**Figure 4 - A Typical X-ray Diffraction Pattern Collected from a Protein Crystal**



This X-ray diffraction pattern was recorded from a crystal of *Erythrina cristagalli* lectin. Data were collected to 2.0Å resolution using the Synchrotron Radiation Source at Daresbury (UK) in December 2001. The diffraction pattern consists of a series of spots (reflections) whose positions contain information on the dimensions of the crystal lattice, and whose intensities are related to the arrangement of the atoms within that lattice.

### 2.3.2 Diffraction by a Crystal

Individual reflections correspond to reciprocal lattice points and are identified by Miller indices (*h*,*k*,*l*). Each diffracted X-ray giving rise to a reflection in the diffraction pattern can be described as the sum of the contributions of all electrons in the unit cell (Rhodes 2000). The value of this sum is known as a structure factor, **F**(*hkl*). The structure factor equation thus represents the molecular transform sampled at reciprocal lattice point *hkl* (Blundell and Johnson 1976):

$$\mathbf{F}(hkl) = \sum_{j=1}^{N} f_j \exp 2\pi i(hx_j + ky_j + lz_j)$$

A crystal will therefore produce a discontinuous diffraction pattern, observed as a series of spots (known as "reflections").

Each reflection has a magnitude |**F**| and a phase angle α. This can be represented as a vector (comprising real and imaginary parts) of length **F** at an angle α to the real axis (Figure 5). The structure factor equation for a given reflection can be expressed in terms of its amplitude and phase:

$$\mathbf{F}(hkl) = F(hkl)\exp i\alpha(hkl)$$

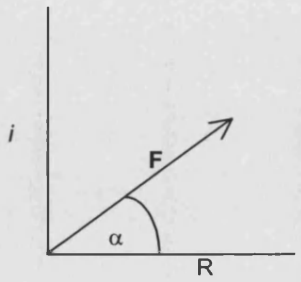
where: *F*(*hkl*) is the amplitude

α(*hkl*) is the phase

The amplitude of the reflection is related to its intensity (*I*) by the equation:

$$I(hkl) = F(hkl)^2$$

**Figure 5 – Vector Representation of a Diffracted Wave**




Vector diagram representing the magnitude (**F**) and phase angle (α) of a recorded reflection.

Mathematically, the recorded diffraction pattern is related to the structure of the protein by a Fourier transform. The Fourier transform of a protein crystal is the product of the Fourier transform of one protein molecule and the Fourier transform of the crystal lattice. In the same way, a Fourier transform of the diffraction pattern will give the electron density of the protein molecule, which can be considered to be an image of its structure. If the structure of the crystal is known, the diffraction pattern can be calculated using the structure factor equation (Blundell and Johnson 1976). In order to compute the structure of the protein from the diffraction pattern, the structure factor equation is rewritten in terms of a continuous summation over the volume of the unit cell:

$$\rho(xyz) = \frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\mathbf{F}(hkl)\exp -2\pi i(hx+ky+lz)$$

where: ρ(*xyz*) is th$\rho(xyz) = \frac{1}{V}\sum_{h=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}\mathbf{F}(hkl)\exp -2\pi i(hx+ky+lz)$

$V$ is the volume of the unit cell

This is known as the electron density equation and it can be seen that if the structure factors for all reflections are known, the electron density at each point in the unit cell may be calculated (Blundell and Johnson 1976).

## 2.4 X-RAY DIFFRACTION DATA COLLECTION

The protein crystal is merely a tool that enables the crystallographer to glean structural information about its component molecules. The crystal lattice acts as a diffraction grating, scattering electromagnetic radiation of appropriate wavelengths (i.e. X-rays) and analysis of the resultant diffraction patterns allows the three-dimensional structures of the molecules that make up the crystal to be determined. It is important to collect good quality diffraction data in order to generate high quality electron density maps and precise atomic models (Dauter 1997).

### 2.4.1 Crystal Handling

The ability of a crystal to diffract cannot be assessed without placing it in an X-ray beam. Some crystals diffract strongly, some weakly and others not at all. This is even true of protein crystals grown in the same drop or under the same conditions in different drops. It may also be difficult to determine the nature of a crystal without recording a diffraction pattern. Mother liquors used to crystallise proteins may contain high concentrations of salt, which can crystallise. Salt crystals are not easy to distinguish from protein crystals by eye, but they produce markedly different diffraction patterns characterised by very few, intense spots.

To collect diffraction data, the crystal must be mounted and centred in an X-ray beam. Protein crystals are sensitive to external conditions and a large proportion of the crystal volume is solvent, so it is important that the crystal does not dry out during the experiment. To collect data at room temperature, crystals are mounted in quartz capillaries with a small volume of mother liquor. Data can also be collected at cryo-temperatures (100-120K) with the crystals mounted in wire loops and positioned under a stream of cold nitrogen gas during the experiment. Cryo-cooled crystals tend to have increased lifetimes in the X-ray beam and produce better diffraction data because cryo-cooling provides some protection against radiation damage (Garman 1999). The most important factor when preparing crystals for cryo-temperatures is to cool them quickly so that ice crystals do not form. A cryoprotectant solution is normally required to act as an antifreeze agent. Crystals can either be crystallised in cryoprotectant mother liquor or immersed in cryoprotectant solution prior to cooling.

Radiation damage is caused by the action of the X-ray beam on the crystal. Energy absorbed by the crystal is dissipated through the crystal in the form of heat and as covalent bond breakage (Sliz et al. 2003). This is "primary" radiation damage and is largely dependent upon the X-ray dose. "Secondary" radiation damage is caused by the action of reactive radical species generated by primary radiation damage within the crystal. This is temperature and time dependent but can be reduced or eliminated by cooling crystals to cryo-temperatures.

### 2.4.2 X-ray Sources and Detectors

Diffraction data can be collected in the "home" laboratory or at a synchrotron radiation source (SRS). The basic requirements are an X-ray source and a detector upon which to record the diffracted X-rays (known as "reflections"). In addition, the experimental set-up will normally include a monochromator, to select a single X-ray wavelength; a mirror, to focus the beam onto the sample; and a positioning device (a goniostat), to allow rotation of the crystal in the X-ray beam.

X-rays of fixed wavelength may be generated in house using a sealed tube or a rotating anode (Drenth 1999). Both produce X-rays by accelerating electrons at high voltage (40-50kV) against a metal target (McRee 1993). The composition of the metal target determines the wavelength at which X-rays are emitted. However, the target becomes very hot and the power of both sources is limited by the rate at which this heat can be dissipated. Rotating anodes are preferred for protein crystallography because the X-rays produced are of higher intensity (i.e. "brighter"), causing less radiation damage to crystals and improving the quality of the recorded data (Nave 1999).

The alternative is a synchrotron radiation source in which electrically charged particles are circulated close to the speed of light using magnetic devices (Drenth 1999). Electrons accelerate as they move in a circle and emit a continuous spectrum of electromagnetic radiation. The X-rays produced by a SRS are up to two orders of magnitude brighter than the best rotating anodes (McRee 1993). Tight monochromation of the X-ray beam reduces the level of background radiation and increases the signal-to-noise ratio. The X-rays are of high intensity, allowing data to be collected from small or weakly diffracting crystals and reducing exposure times. However, if the intensity of the radiation is too high, the level of radiation damage in the crystal will increase (Nave 1999). The X-ray beam is also highly parallel with low divergence, giving rise to sharp diffraction spots. The maximum achievable resolution using a SRS is often better than that in house because there are more diffracted

beams at high diffraction angles. The wavelength of the X-rays may also be varied (i.e. it is tunable), which affords synchrotron radiation sources two additional advantages over rotating anodes and sealed tubes: Shorter wavelength X-rays are absorbed to a lesser extent by crystals, causing less radiation damage (Helliwell 1997) and the tunability enables multi-wavelength anomalous dispersion experiments to be performed (Walsh et al. 1999).

Diffraction experiments generally require X-rays of a single wavelength, so a narrow band of radiation is selected from the spectrum output by the source (Drenth 1999). The most commonly used radiation from an in-house source, CuKα, has a characteristic wavelength of 1.5418Å. This wavelength is selected from the spectrum produced by a copper anode using a nickel filter or a graphite monochromator (Drenth 1999). Monochromators at synchrotron radiation sources are generally composed of silicum or germanium, which allow selection of a narrower wavelength band. The X-ray beam can be focussed using a pair of curved mirrors, one mirror in the horizontal plane and one in the vertical plane (McRee 1993). Focussing increases the brilliance of the radiation incident on the crystal and allows resolution of large unit cells. The direct beam is extremely powerful and is blocked by a piece of metal (the beamstop) placed just between the crystal and the detector. It prevents excessive amounts of radiation reaching the centre of the detector.

The detector records the positions and intensities of the X-rays diffracted by the crystal. It must therefore be uniformly sensitive and record a large proportion of the incident X-rays with high efficiency (Amemiya 1997). The diffraction pattern may be distorted if the detector has a flat surface because X-rays are emitted from the crystal in a sphere. Distortion can be recognised by the presence of elongated spots at the edges of the diffraction pattern and must be corrected. Diffraction data were originally recorded using photographic film, which has a high resolution and can be bent to form a curved surface (McRee 1993). However, film records a high level of background radiation, has low efficiency (a small number of the incident waves are recorded) and a low dynamic range (i.e. the reflections are easily saturated). The level of background radiation is not negligible because there is significant scattering of X-rays by air (Drenth 1999). Exposed X-ray film must then be developed and scanned in order to analyse the recorded data but, in order to retain the high resolution, a fine pixel scanner must be used. Most beamlines now employ image plates and charge-coupled device area detectors in place of X-ray film.

Image plates comprise a thin layer of an inorganic phosphor deposited on a flat base (Drenth 1999). X-rays incident on the image plate excite electrons in the

phosphor to higher energy levels. Most of the energy is stored in the phosphor, but some is emitted as fluorescent light. The stored energy is released by illuminating the plate with a red laser and the amount of blue light emitted is proportional to the number of incident photons. The diffraction pattern is read spirally but the read-out time is too slow for image plates to be used with synchrotrons. Once an image has been recorded and read, it can be erased by exposure to intense white light and the image plate used to record another diffraction pattern. Image plates are extremely sensitive at shorter wavelengths and are much more sensitive than X-ray film (Drenth 1999). The main problem is with "overloaded" reflections. If a pixel becomes overloaded, the statistical error will be artificially low and such reflections should be removed from the data. Charge-coupled device (CCD) area detectors utilise inorganic phosphors that emit visible light in response to X-radiation. The CCD accumulates charge in direct proportion to the number of photons incident upon it and the charges are read out serially (Rhodes 2000). CCDs are suitable for use with synchrotrons because they have better spatial resolution and shorter read-out times than image plates (Muchmore 1999). Other advantages include high dynamic range with low noise and high maximum count rate, although they have lower sensitivity at shorter wavelengths.

### 2.4.3 Data Collection Strategy

The success of subsequent steps in protein structure determination depends largely upon the quality of the diffraction data recorded from the crystal (Dauter 1999). The detector records the positions and intensities of X-rays diffracted by the crystal and it is important that these reflections are recorded completely and do not overlap. As they form the only experimentally observed parameter, the quality of the intensities is important. The whole area of the detector face should be used to record the diffraction pattern so the crystal-to-detector distance should be as large as possible to collect data to the resolution limit of the crystal (Dauter 1997). A large crystal-to-detector distance also increases the signal to noise ratio. As the resolution increases, so does the quantity of data. The exposure time per image should also be carefully chosen – if it is too long, the detector will be overloaded with strong, low-angle reflections and the level of radiation damage to the crystal will increase (Evans 1999).

Reducing the level of background radiation recorded by the detector increases the signal to noise ratio and improves the quality of the data. This can be achieved by matching the width of the X-ray beam to the size of the crystal (Evans 1999). A large beam incident on a small crystal will produce a high level of background radiation because X-rays are scattered by air. The path length of the direct X-ray beam through

air should therefore be minimised by positioning the beamstop close to the crystal (McRee 1993).

In order to collect a complete dataset, it is necessary to rotate the reciprocal lattice so that all reciprocal lattice points will, at some time, lie on the Ewald sphere. This is achieved by rotating the crystal during data collection (because the reciprocal lattice rotates with the crystal lattice). A dataset therefore comprises a set of X-ray diffraction images recorded across a rotation range in a series of small oscillations. The oscillation angle should be as large as possible to minimise the number of images that must be recorded while avoiding overlapping reflections. The number of images required for a complete dataset and high redundancy depends on both crystal symmetry and the angular range which each image covers. High redundancy is important because it improves the accuracy of the data (Evans 1999).

The recorded reflections will be small and sharp for a highly ordered crystal (McRee 1993). However, most crystals contain imperfections and, in some cases, the reflecting planes diverge increasing the size of the reciprocal lattice points and hence the angular range over which they diffract. This "mosaicity" can be recognised in the diffraction pattern by wide lunes and overlapping reflections. High mosaicity produces lower-resolution diffraction and a lower signal-to-noise ratio. Overlaps occur when two reciprocal lattice points meet the Ewald sphere at the same point and within one oscillation of the crystal, giving rise to reflections in the same place in the diffraction pattern, making it difficult to deduce from which reflective plane the diffracted beams originate.

There is a "blind region" in the part of reciprocal space close to the axis of rotation where the reciprocal lattice points will never meet the surface of the Ewald sphere (Dauter 1997). For higher symmetry space groups, it can be possible to orient the crystal so that reflections equivalent to those in the blind region can be measured.

## 2.5 DATA PROCESSING

In order to reconstruct the three-dimensional structure of a protein from a set of diffraction patterns, the amplitudes and phases of the reflections must be known. The amplitude of each reflection can easily be determined because it is related to the recorded intensity. Therefore, the diffraction data must be carefully processed. Data processing, the analysis and reduction of diffraction data, comprises several steps (Otwinowski and Minor 1997): Firstly, the recorded diffraction images are visualised and the reflections indexed. The crystal and detector parameters must then be refined. The intensities of the reflections are integrated and the relative scale factor between

measurements is determined. Crystal parameters are precisely refined using the whole dataset and measurements related by space group symmetry are merged and statistically analysed.

### 2.5.1 Indexing

The first step in data processing is to give each reflection an index that can be used to identify each point in reciprocal space (McRee 1993). Indexing begins with the identification of intense spots on the diffraction pattern – a process known as "peak picking" (Rossmann and van Beek 1999). The positions of these peaks are analysed in order to elucidate the unit cell dimensions and Bravais lattice of the crystal. The symmetry of the crystal is then selected from a list that ranks each Bravais lattice with a "penalty". In general, the correct lattice is that with the highest symmetry and lowest penalty ranking. It is worthwhile to note that any unit cell dimensions will fit a triclinic lattice, but that a higher symmetry lattice should be sought before a P1 space group is assigned. The orientation of the crystal with respect to the direction of the X-ray beam must also be determined, requiring accurate knowledge of the main beam position.

Once the crystal orientation is known and the space group has been chosen, Miller indices (*hkl*) are assigned to each reflection. The positions of all the reflections in the diffraction pattern can then be predicted for each image and differences between the positions of the observed and calculated reflections are minimised by refinement (Pflugrath 1997). Weak reflections can only be located by prediction based on the positional information provided by stronger reflections. The orientation of the crystal and detector are refined for each image, and this is important for refining the positions of the reflections. Accurate prediction of reflection positions is essential for precise integration of intensities. Errors in the prediction of spot positions affects the statistical error in the intensity measurements (Otwinowski and Minor 1997).

### 2.5.2 Integration

Integration of the intensity (I) of a spot in the diffraction pattern separates the background counts from those in the reflection (McRee 1993). However, in order to accurately integrate the intensities of the reflections, the spots in the diffraction images must be non-overlapping and the index and position of each reflection must be known (Otwinowski and Minor 1997).

Two methods are available for the integration of reflection intensities, summation and profile fitting, and there should be good agreement between the results that each method produces for any given dataset. Summation involves summing the

intensities of all the pixels in a peak, then subtracting the intensity of the background radiation (Leslie 1999). The intensity of the background radiation is determined using the pixels adjacent to a peak and a background-corrected intensity is calculated for each reflection. Profile fitting involves fitting a smooth profile to the data and taking the area beneath the curve as the intensity (McRee 1993). This has two advantages over summation – an improved signal-to-noise ratio for weak reflections and estimation of overloaded reflections (Leslie 1999). Firstly, a standard profile accurately representing the true reflection profile is derived, with the optimum profile being that which provides the best fit to all the reflections. From this standard profile, the intensities for fully recorded reflections can be evaluated by determining a scale factor and background-plane constants. For both methods, it is important to accurately estimate the level of background radiation recorded by the detector in order to reduce the error in the measured intensity of each reflection.

Integrated intensities are merged in order to group together reflections with common indices (McRee 1993). Merged multiple measurements of several reflections are weighted by the estimated error ($\sigma$). Reflections are often rejected during data processing by the ratio of intensity to the error – $I/\sigma(I)$ – which, for a good dataset, should have a value greater than 2. The error ($\sigma$) estimated for each reflection combines the height and variance of the background intensity and the number of times each reflection was measured. Errors in intensity measurement arise from several sources, including counting statistics; instability of the detector; profile fitting; local variations in the level of background radiation; saturated pixels; overlapped profiles; and errors in background models (McRee 1993). For example, the diffraction pattern may include features (such as white radiation streaks and cosmic rays) that do not arise from Bragg diffraction and the effects of these on the estimation of the background intensity must be reduced (Leslie 1999).

### 2.5.3 Scaling

Scaling corrects for absorption differences, variation in the intensity of the X-ray beam and, if more than one crystal is used, variability in the diffracting power of crystals during data collection (Drenth 1999). If diffraction data are collected from more than one crystal, the separate datasets must be normalised by the application of a scale factor before they can be merged (Monaco 1994). The dataset is then reduced to a unique list of reflections (i.e. "merged").

Scaling generates statistics that evaluate the quality of the data and should be performed several times to improve these statistics as outliers are rejected. An R factor

($R_{symm}$) compares symmetry-related reflections that should be of identical intensity and, thus, provides an internal measure of the accuracy of a dataset (McRee 1993).

$$R_{symm} = \frac{\sum_h (I_h - \bar{I}_h)}{\sum \bar{I}_h}$$

If $R_{symm}$ has a value of 0.10 or less, the data are considered good enough for structure determination. A $\chi^2$ test performed on the scaled data indicates the accuracy of the error model. A good error model will give a $\chi^2$ value close to 1.0. A high value of $\chi^2$ may indicate that the crystal suffered radiation damage, which cannot be corrected by scaling. The completeness of the data is given as a percentage and is the number of measured unique reflections compared with the number of possible unique reflections. Generally, data that are >80% complete through all resolution bins are considered sufficient for accurate structure determination.

Scaling programs can also be used to determine the space group of a crystal. The Bravais lattice is determined during indexing but some lattices include several spacegroups. Scaling is performed on each possible space group, starting with that of lowest symmetry. If the value of $\chi^2$ increases with higher symmetry, then the previous choice of space group was correct. Scaling produces a list of systematic absences – if the correct space group is chosen, these reflections should be absent and their values very small. Correspondence of absences observed in the experimental data with the systematic absences characteristic of certain space group, identifies the correct space group for the experimental data.

## 2.6 THE PHASE PROBLEM

As described previously, the electron density in a crystal can be determined by performing a Fourier summation:

$$\rho(xyz) = \frac{1}{V} \sum_{hkl} |F(hkl)| \exp[-2\pi i(hx + ky + lz) + i\alpha(hkl)]$$

where: $|F(hkl)|$ is the structure factor amplitude of reflection hkl

$\alpha(hkl)$ is the phase angle

$x,y,z$ are coordinates in the unit cell

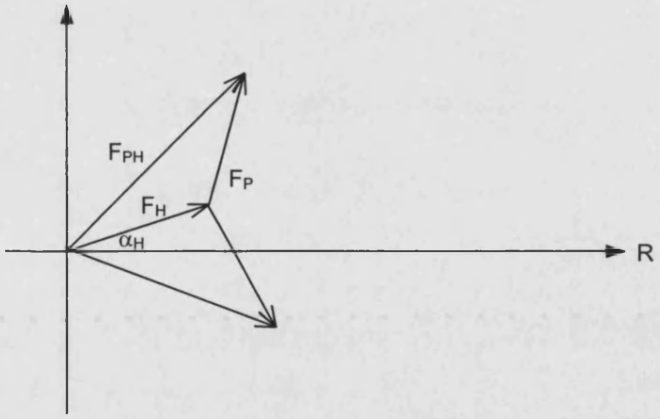
The amplitude of each reflection can easily be determined from the measured intensity, because $I(hkl) = F(hkl)^2$, but the phase angles cannot be measured experimentally. This creates a problem, known as the "phase problem", because the phases contain most of the structural information. It follows that accurate phase information is essential for the generation of good quality electron density maps.

Initial phases, to be used for the generation of an initial electron density map, must therefore be derived indirectly and there are three main methods are available for this. Isomorphous replacement uses differences in intensity between the diffraction patterns of a native protein crystal and at least one heavy atom protein derivative. Multi-wavelength anomalous dispersion derives initial phases from the anomalous scattering of X-rays by heavy atoms present in the crystal at wavelengths close to their absorption edges. Molecular replacement utilises phase information from a previously determined, homologous protein structure.

### 2.6.1 Isomorphous Replacement

The isomorphous replacement method (Green et al. 1954) involves the preparation of heavy atom derivatives either by soaking protein crystals in a buffer containing a heavy metal compound or by co-crystallising a protein with a heavy metal compound (Ke 1997). The presence of the heavy atoms (such as Hg or Pt) alters the intensities of reflections in the diffraction patterns. These differences can be used to locate the positions of the heavy atoms in the crystal and determine the relative phase angles.

**Figure 6 - Phase Determination by Isomorphous Replacement**




The magnitudes of $|F_{PH}|$ and $|F_P|$ are measured experimentally using the derivative and native crystals. Values for $|F_H|$ and the phase of the heavy atom are determined from the heavy atom model. This gives two possible solutions for the structure factor of the native protein. A unique solution can only be obtained by collecting data from a second heavy-atom derivative because this produces another pair of vectors that resolve the phase ambiguity.

The structure factors of the native crystal ($\mathbf{F}_P$), the heavy atom ($\mathbf{F}_H$) and the heavy atom derivative ($\mathbf{F}_{PH}$) are related as follows:

$$\mathbf{F}_P = \mathbf{F}_{PH} - \mathbf{F}_H$$

The magnitudes of $|F_P|$ and $|F_{PH}|$ are measured experimentally and values for $|F_H|$ and $\alpha_H$ are obtained from the heavy atom model. Representing the structure factors as vectors reveals two possible phase angles for $\mathbf{F_P}$ (Figure 6), one of which is correct. A unique solution can only be obtained by collecting data from a second heavy-atom derivative as this produces another pair of equations that resolve the ambiguity.

The use of only one heavy atom derivative is known as single isomorphous replacement (SIR) but will not solve the phase ambiguity for the native protein. The use of two or more derivatives (multiple isomorphous replacement, MIR) will solve the phase ambiguity.

### 2.6.2 Multiwavelength Anomalous Dispersion

The multiwavelength anomalous dispersion (MAD) method (Hendrickson and Ogata 1997) requires tunability of the X-ray wavelength, so the experiments can only be performed using a SRS. This is because the technique relies upon the anomalous scattering produced when a bound electron is promoted to a higher energy level. The energy required comes from photon absorption at the "absorption edge" of the element. The energy, E, imparted to the electron is related to the X-ray wavelength, $\lambda$, as follows:

$$E = h\frac{c}{\lambda}$$

where: h is Planck's constant

c is the speed of the incident radiation

X-rays used in standard diffraction experiments do not cause electronic transitions in the "light" elements that make up proteins (i.e. C, H, N and O). Anomalous scattering is only observed when heavy atoms are present in the protein and the wavelength of the radiation is varied to bring the energy close to that required for an electronic transition. Heavy atoms, such as zinc and iron, are intrinsic to some proteins. Other proteins can be recombinantly expressed such that all methionines are replaced with selenomethionines, which are no larger than the native amino acid and do not affect the structure of the protein.

As the energy of the X-rays incident on a crystal approaches that required for an electronic transition, the oscillation of the electron is increased and a resonance condition is established (Hendrickson and Ogata 1997). Remote from the resonant energy of the electron, scattering is largely independent of wavelength and produces a primary wave 180° out of phase with the incident radiation. Close to resonant energies resonance effects cause anomalous scattering. The effect of anomalous dispersion can

be expressed in terms of the atomic scattering factor, $f$, which may be defined as follows (Karle 1989):

$$f = f_0 + \Delta f + if''$$

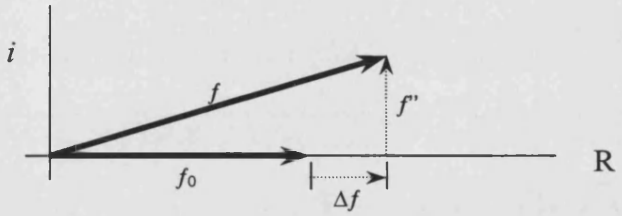
where: $f_0$ is the normal or non-anomalous scattering factor

$\Delta f$ is the real part of the correction to $f_0$

$if''$ if the imaginary part of the correction to $f_0$

This is explained in Figure 7. When the value of $f''$ is not negligible, Friedel's law (which states that $\mathbf{F}_{hkl} = \mathbf{F}_{-h-k-l}$) does not hold true. The difference between the structure factors of two Friedel mates, the Bijvoet difference, is used to determine phase information (Blundell and Johnson 1976).

**Figure 7 - The Anomalous Scattering Factor**




The atomic scattering factor for a completely free electron is represented as $f_0$. The anomalous scattering factor for a heavy atom ($f$) consists of two additional components, a real part $\Delta f$ and an imaginary part $f''$.

The X-ray wavelengths used in a MAD experiment must be carefully chosen to maximise the Bijvoet difference (Drenth 1999). Data are typically recorded at three different wavelengths – where $f'$ is small, where $f'$ is large and negative and where the Bijvoet difference is maximum. Locating the anomalous scatterers in the structure can be achieved in a manner analogous to that used in isomorphous replacement, enabling the structure factor amplitudes of the protein to be estimated.

### 2.6.3 Molecular Replacement

Proteins that are homologous in their amino acid sequences often have similar tertiary structures. Therefore, if the protein under investigation is similar in primary structure to another protein whose crystal structure has previously been determined, the phases of the known structure can be used as initial phases for the unknown protein. This is known as molecular replacement (Rossmann and Blow 1967). However, if a low homology model is used, the phases will poorly represent the true phase of the unknown protein and the resultant structure will be biased towards the

model and consequently difficult to refine correctly (McRee 1993). The bias introduced by the model phases is the main disadvantage of the molecular replacement method. For this reason, there should be high structural homology betweenthe known and unknown proteins. Suitable model proteins can be identified by comparing the sequence of the unknown protein with the sequences of proteins whose structures have been deposited in the PDB.

Two identical molecules ($X_1$ and $X_2$) with different positions and orientations can be brought to coincidence by defined rotation and translation functions (Blundell and Johnson 1976) as follows:

$$x_2 = [C]X_1 + d$$

where: [C] is the rotation matrix

d is the translation vector

Thus, molecular replacement identifies the rotation and translation that allow a known structure (the "search model") to be precisely positioned in the unit cell of the unknown structure (the "target"). It is a six-dimensional search problem because three rotational and three translational parameters are involved. If the asymmetric unit contains more than one protein molecule or subunit, molecular replacement can also determine their relative positions (i.e. non-crystallographic symmetry) (Drenth 1999).

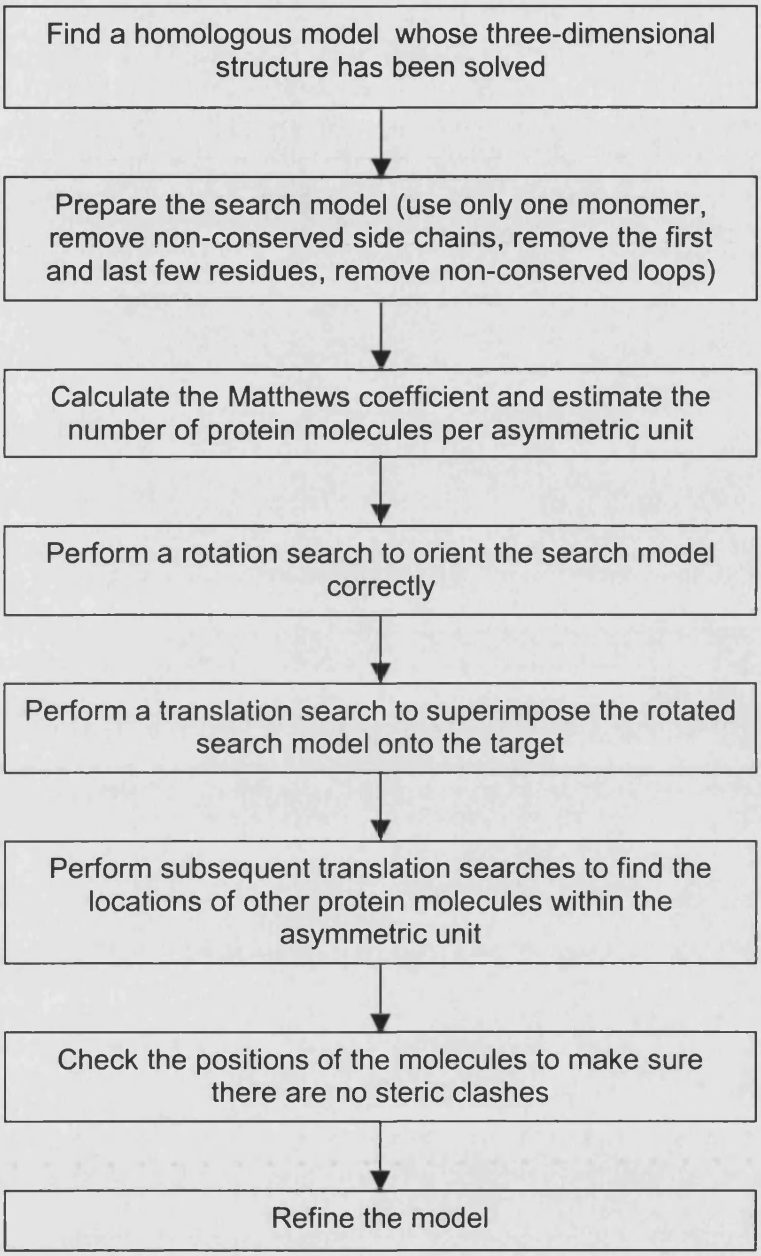
Figure 8 is a flowchart of the steps involved in molecular replacement. First, a model structure must be identified and modified, if necessary, to improve the success of the rotation search. The number of molecules per asymmetric unit is estimated and the rotational and translational parameters required to superimpose the search model on the target are ascertained. The solution is checked to ensure that no molecules are overlapping in the unit cell and the model structure is then refined.

Solving the rotation and translation functions is not always clear-cut (Drenth 1999). It can be necessary to vary the parameters of the search or modify the search model. For example, if the model is an oligomer it may be better to perform the search with only one monomer. It may also be beneficial to remove the first and last few amino acids from the termini; to delete non-conserved loops; and/or to replace non-conserved residues with alanines. These alterations reduce the difference between the search and target structures and improve the success of the rotation and translation functions.

Before molecular replacement is undertaken, the number of protein molecules per asymmetric unit should be estimated by calculating the Matthews coefficient, Vm (Matthews 1968). This provides an estimate of the crystal solvent content from the unit cell volume and molecular weight of one molecule. Most protein crystals are 30-75%

protein, corresponding to Vm values of 4.00 to 1.66. If the value of Vm falls outside this range, the assumed number of molecules per asymmetric unit is likely to be incorrect.

**Figure 8 - Flowchart of Steps Involved in Molecular Replacement**




The molecular replacement method involves two main steps, a rotation search and a translation search, that allow the search model to be precisely positioned in the target unit cell. It is actually a six-dimensional problem because it involves three rotational and three translational parameters. It is important to choose a highly structurally homologous model because the initial phases will be biased towards the search model.

### 2.6.3.1 The Patterson Function

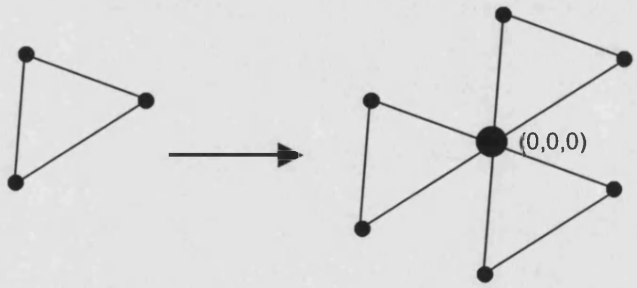
The basic principle of molecular replacement can be explained in terms of the Patterson function, *P*(*uvw*). This is a Fourier summation involving the measured intensities, with all phase angles set to zero (Drenth 1999):

$$P(uvw)=\frac{1}{V}\sum_{hkl}|F(hkl)|^2\cos[2\pi(hu+kv+lw)]$$

The Patterson cell has the same dimensions as the real unit cell of the crystal, but the coordinates u,v,w are used to avoid confusion with the coordinates of the real cell (x,y,z).

The Patterson synthesis produces a map of the vectors between atoms in the unit cell by placing each atom in turn at the origin (Figure 9). If the unit cell contains N atoms, the Patterson map will have $N^2$ peaks, of which N will be located at the origin (Drenth 1999). This explains why the highest peak is the map is located at the origin. The number of unique peaks in the map is therefore N(N-1). The Patterson map will always be centrosymmetric.

**Figure 9 - Vectors in a Patterson Map**




The Patterson synthesis of a three-atom molecule (shown on the left) is shown on the right. The Patterson map shows the vectors between the atoms and is derived by taking each atom in the molecule in turn and placing it at the origin (0,0,0).

Patterson vectors relate pairs of atoms in the unit cell and the peaks in the Patterson map correspond to these interatomic vectors (Grosse-Kunstleve and Adams 2001). Self-Patterson vectors are relatively short and exist between atoms from the same protein molecule, whereas cross-Patterson vectors relate atoms from different protein molecules. If more than one identical molecule exists in the asymmetric unit, the distribution of self-vectors will be the same for all of the molecules except for a

rotation that is equal to the non-crystallographic symmetry. In molecular replacement, both self- and cross-Patterson vectors are used to superimpose the search model onto the target in the target unit cell.

Self-Patterson peaks are located around the origin of the Patterson map in a volume of radius equal to the dimension of one molecule (Drenth 1999). The self-vectors for homologous molecules are very similar, but not equal, and can reveal the rotation that relates them. Cross-vectors can then be used to determine the translation required to move the rotated search model to its correct position in the target unit cell. An observed Patterson map, calculated from the measured reflection intensities, and a model Patterson map, derived from the search model after rotation and translation, can be compared by superposition (Grosse-Kunstleve and Adams 2001). The observed and model Patterson maps will overlap maximally if the search model has been correctly oriented.

### 2.6.3.2 Rotation Function

The rotation function determines a rotation matrix, [C], that transforms the coordinates representing the search model without altering its dimensions. The search model, which is assumed to be spherical, is placed at the centre of an artificial unit cell with no crystallographic symmetry (i.e. space group P1) (Drenth 1999). Calculated structure factors for the search model in this artificial lattice can then be used instead of X-ray data from the known crystal structure in the rotation search. The rotation search determines the spatial orientation of the search model with respect to the target and is performed over the unique volume of the space group of the unknown structure. The search model is rotated through all possible angles and, at each point, the observed and model Patterson maps are compared (McRee 1993). If the search model is correctly oriented, there will be maximum overlap between the two maps.

Only intramolecular vectors, which are relatively short, should be considered during the rotation search, so the dimensions of the unit cell are limited to eliminate the longer cross-vectors. Typically, the cell will be no more than 30Å larger than the search model in each dimension. The resolution of the data is another important consideration. Very low-resolution data are excluded because they only supply information on the gross features of the structure. Very high-resolution data should also be removed, as they will be too-heavily biased towards the search model. Therefore, the resolution range used for a rotation search is usually limited to 10-3.5Å (McRee 1993).

### 2.6.3.3 Translation Function

After the rotation angles have been determined, the translation required to superimpose the search model onto the target must be ascertained (Drenth 1999). This is achieved by moving the search model through the asymmetric unit. Observed and calculated structure factors ($F_{obs}$ and $F_{calc}$) are then compared at each position by calculating an R-factor or correlation coefficient. The correlation coefficient is a more reliable indication of correct placement of the search model because it is insensitive to scaling. If there is more than one molecule or subunit per asymmetric unit, then the translation function must be performed repeatedly until all of the molecules have been located. Possible locations of the search model are limited by knowledge of the crystal packing because the protein molecules cannot penetrate each other. However, the translation vector cannot be successfully assigned if the rotation matrix has not been precisely determined (Blundell and Johnson 1976).

## 2.7 REFINEMENT AND MODEL BUILDING

Unfortunately, the methods used for phase determination are not free of error and, consequently, interpretation of the resultant electron density maps is not always straightforward. The level of detail observed in the electron density map depends upon the quality of the diffraction data, so it is vital that a complete dataset with a high signal-to-noise ratio and high redundancy is used (Jones and Kjeldgaard 1997). The initial molecular model must be built into the calculated electron density. The ultimate aim of model building and refinement procedures is to generate a model of the protein that accurately represents the experimental data and makes sense physically, chemically and biologically (Kleywegt and Jones 1997). The processes involved in model building and refinement are repeated until no further improvements can be made to the model. The structure of the protein can then be analysed to extract information on its function or mechanism of action, for example.

### 2.7.1 Refinement

Refinement is achieved by modifying the initial molecular model to improve the agreement between the observed and calculated structure factor amplitudes (Kleywegt and Jones 1997). The agreement between $F_{obs}$ and $F_{calc}$ is represented by the R-factor, $R_{cryst}$ (Drenth 1999):

$$R_{cryst} = \frac{\sum_{hkl} \left| |F_{obs}| - k|F_{calc}| \right|}{\sum_{hkl} |F_{obs}|}$$

A high R-factor indicates that the model is incorrect (McRee 1993) but, because refinement is not a fully automated process, it is possible to over-fit the model, giving rise to an artificially low R-factor. So, the value of $R_{cryst}$ alone is not a good indication of the accuracy of the model. Refinement can be monitored using another R-factor known as $R_{free}$, which is calculated in the same way as $R_{cryst}$ from a small subset of the data (the "test set") that are excluded from the refinement. Comparison of $R_{cryst}$ and $R_{free}$ provides a more reliable indication that the model is being improved – if the model forms an accurate representation of the experimental data, both R-factors will have similar values (Kleywegt and Jones 1997). Once the R-factors have reached a minimum value and are in agreement, one last round of refinement is performed before the accuracy of the final model is assessed.

Refinement is concerned with the coordinates (*x*,*y*,*z*) and B-factors of all the atoms within the model. The B-factor measures the displacement of an atom due to thermal motion, conformational disorder and static lattice disorder (McRee 1993). This displacement causes the electron density to spread, making the maps more difficult to interpret. The number of measured reflections in the dataset is far outweighed by the number of parameters to be refined. Therefore, additional information is included in the refinement process. This includes information on ideal bond lengths and bond angles, the amino acid sequence of the protein and "solvent flattening". Refinement is described as "restrained" if the stereochemical parameters are allowed to vary around a set standard or, "constrained" if the stereochemistry of bond lengths and angles is kept rigid and only dihedral angles can be varied (Drenth 1999). Solvent flattening prevents ordering of the bulk solvent so it is observed as a flat region in the electron density map.

### 2.7.2 Calculation of Electron Density Maps and Model Building

As has been described previously, just as the diffraction pattern is a Fourier transform of the protein and the crystal lattice, a Fourier transform of the diffraction pattern with phase information will provide an electron density map of the protein.

The classic Fourier synthesis involves the observed amplitudes ($F_O$) and the most current phase angles ($\alpha_{calc}$) but is subject to model bias (McRee 1993). A difference Fourier map ($F_O$-$F_C$), calculated using observed amplitudes, calculated amplitudes ($F_C$) and current phases, shows peaks where electron density is not accounted for in the model. Thus, it is used for identifying corrections that should be made to the model. The $2F_O$-$F_C$ map (the sum of the sum of the $F_O$ map and the $F_O$-$F_C$

map) looks like protein density, but its quality is dependent upon the quality of the phases.

$2F_O$-$F_C$ and $F_O$-$F_C$ electron density maps are compared to aid model building. The original maps should be retained in order that they may be reinterpreted after each round of refinement (Kleywegt and Jones 1997). As well as protein molecules, the model may also contain metal ions, ligands, substrates and water molecules, which must be modelled into the $F_O$-$F_C$ electron density. If molecular replacement was used to determine the initial phases, it may also be necessary to replace non-identical side-chains that were converted to alanine in the search model back to the native side-chain in the structural model. Such residues will need to have their native side-chains rebuilt in accordance with the $F_O$-$F_C$ electron density map. After an amino acid residue has been built or modified, or an additional molecule added to the model, the model should be refined. New electron density maps should be generated and the R-factors analysed to determine that the change led to an improvement in the model.

Towards the end of refinement and model building, $F_O$-$F_C$ electron density can often be explained by the addition of water molecules to the molecular model (McRee 1993). The structure of the solvent surrounding the protein molecules is an important aspect of the protein structure and should be modelled carefully. Water molecules should not be added until all other changes have been implemented and the model is almost complete because not all unexplained density is due to waters (McRee 1993). For example, buffer molecules and/or salt present in the mother liquor may be bound by the protein and should be modelled as such in the electron density map. Correctly identifying peaks in the $F_O$-$F_C$ map will lead to a reduction in $R_{cryst}$ and $R_{free}$. Water molecules should be placed in areas of good $F_O$-$F_C$ electron density where they may make hydrogen bonds with suitable residues in the model (Kleywegt and Jones 1997). Waters close to the protein surface typically form hydrogen bonds with polar groups on the protein. The total number of waters that can be added depends upon the maximum resolution and the quality of the maps but, for a good model, two waters can be modelled for every amino acid residue.

### 2.7.3 Structure Validation

The crystallographic map is a time-average of all the molecular structures in the crystal and, therefore, may not represent a true structure, especially if the protein is in motion. So, the final model should be validated before it is analysed.

The error in the atomic coordinates of the molecular model is calculated using the Luzzati plot. This analysis is based upon the assumption that the difference

between $|F_{obs}|$ and $|F_{calc}|$, as quantified by the R-factor, is due solely to errors in the positional coordinates of the atoms (Drenth 1999). For a well-refined model, the Luzzati coordinate error can be expected to fall in the range 0.2-0.3Å. B-factors are also calculated for each non-hydrogen residue in the model (McRee 1993). The average B-factors in a well-ordered model should not exceed 30Å$^2$, although the values for water molecules and surface loops may be slightly higher (Drenth 1999).

Structural analysis of peptides has provided a wealth of information on the stereochemistry of protein functional groups and these data can be used to evaluate the quality of a crystallographically-determined protein structure (McRee 1993). Ideal values for bond distances, bond angles, dihedral angles and van der Waals contact distances are compared with the molecular model to calculate its deviation from ideal stereochemistry (Kleywegt and Jones 1997). Bond lengths and bond angles should not deviate from ideal values by more than 0.03-0.05Å or 3°, respectively. The main chain dihedral angles ($\phi$ and $\psi$) in a protein can only take on a limited range of values and the Ramachandran plot (Ramachandran et al, 1963) highlights those values that are stereochemically viable. The plot is divided into four regions (allowed, additional allowed, generously allowed and disallowed) and a good structural model will have at least 90% of amino acid residues in the core regions of the plot (Morris et al. 1992). If a residue does not meet these criteria, the structure should be analysed to uncover any features that may explain the deviation.

The shape and chemical characteristics of the protein surface determine the nature of its interactions with other biological molecules (McRee 1993). So, structural analysis is concerned with studying the structure of the protein to glean information on its function and to uncover any unique features. Lattice contacts should be checked to find protein-protein interactions, such as those occurring at the interfaces of multimeric proteins. Lattice contacts also provide information on the packing of molecules in the crystal lattice – if atoms are closer than van der Waals packing distances allow, there is some error in the model because two atoms cannot occupy the same position in the crystal (McRee 1993). Information on the formation of hydrogen bonds with other protein molecules, ligands and the solvent is also useful for understanding the function and overall structure of the protein. A hydrogen bond is assumed to be formed if a proper acceptor-donor pair exists within the correct bonding distance of 2.7-3.0Å (McRee 1993). If the protein under investigation is an enzyme, information on the protonation state of certain amino acids can also provide clues as to the mechanism of action. The surface of the protein is studied to determine the solvent accessibility of specific residues and to look for features, such as depressions or clefts that may

indicate the location of a binding or active site. Interactions between the protein and any bound ligands should be carefully investigated in order to glean information on the function and/or mechanism of action of the protein.

The final step in protein crystallography is deposition of the coordinates of the protein in the PDB to enable other scientists to access the information. There is an obligation to deposit atomic coordinates within a certain period after the structure has been solved and published.

# CHAPTER THREE
# CRYSTAL STRUCTURE OF *ERYTHRINA CRISTAGALLI* LECTIN

## 3.1 INTRODUCTION

Lectins are carbohydrate-binding proteins of non-immune origin. They do not exhibit catalytic activity and thus cannot be described as enzymes. The carbohydrate-binding site is therefore not an active site, and is known as a "combining site". Lectins of different families exhibit variation in their three-dimensional structures (reviewed in Loris 2002) and exist in a variety of oligomeric states. Lectins are also found in a wide range of species, from bacteria and viruses to plants and animals.

Lectins recognise and bind to a number of mono-, di- and oligosaccharides as well as to glycoproteins to mediate a number of cell-cell interactions and biological processes. Protein-carbohydrate interactions represent an important mechanism of cellular targeting and recognition. Oligosaccharides post-translationally added to many proteins act as a recognition motif and are, therefore, the most likely natural ligands of lectins.

### 3.1.1 Glycosylation of Proteins

Glycosylation of proteins is the most common form of post-translational modification in eukaryotes, particularly among membrane proteins and secretory proteins (Ashford and Platt 1999). The addition of saccharide units to proteins not only affects their physical properties, such as solubility, but also influences their activities.

#### 3.1.1.1 N-Linked Glycosylation

Sugars are commonly linked to eukaryotic proteins by the formation of an N-glycosidic bond between N-acetylglucosamine (GlcNAc) and the side chain of an asparagine residue. The target asparagine must be located within the motif Asn-X-Thr/Ser (where X can be any amino acid except proline). N-glycosylation of proteins begins with the addition of a dolichol-linked "high-mannose" structure, comprising two GlcNAc, nine mannose and three glucose residues. This oligomannose saccharide is assembled in the endoplasmic reticulum linked to dolichyl pyrophosphate and transferred to the target protein by an oligosaccharyltransferase enzyme. The high-mannose structure can then be further processed to produce a variety of oligosaccharide units (Spiro 2000; Turco and Robbins 1979).

### 3.1.1.2 Role of Glycosylation

The post-translational addition of sugars to proteins provides a mechanism by which other proteins can specifically recognise them. In addition to the most common N-linked form of glycosylation, sugars can be linked to proteins by O-glycosidic linkages, C-mannosylation, phosphoglycation and glypiation (reviewed in Spiro 2002). O-linked glycosylation attaches sugars to amino acids with side chains that contain hydroxyl groups (i.e. serine, threonine, tyrosine, hydroxyproline and hydroxylysine). C-mannosyl bonds, first observed in Eosinophil Derived Neurotoxin, can be formed between an α-mannosyl residue and C2 atom of a tryptophan residue (Krieg et al. 1997). The formation of a phosphodiester bond between a sugar and a protein is known as phosphoglycation and involves serine residues. Glypiation is the linkage of a glycophosphatidylinositol anchor to proteins by the formation of a bond between the target protein and phosphoethanolamine linked to mannose.

One common role of the bound carbohydrate is to affect the rate of clearance of proteins from serum. Oligosaccharides are also recognition molecules that are involved in interactions with a variety of cells and biological molecules. Glycosylation can target proteins to specific regions of the cell membrane or intracellular compartments and defects in glycosylation pathways can therefore result in disease. Lysosomal enzymes are targeted to the lysosome from the Golgi apparatus by the addition of mannose-6-phosphate (Griffiths et al. 1988; Kornfeld and Mellman 1989), which is recognised by a specific lectin receptor (Dahms et al. 1989). I-cell disease is caused by defects in the enzymes that attach mannose-6-phosphate to proteins destined for lysosomes.

### 3.1.1.3 Cross-Linking and Multivalency

Lectins exhibit high specificity for oligosaccharides expressed on cell surfaces, binding their target ligands with high affinity. This high level of specificity and affinity (known as avidity) can be achieved in several ways: the presence of several recognition sites on ligands (i.e. ligand multivalency); the presence of an extended combining site capable of accommodating several sugar residues; and/or the clustering of several binding sites by oligomerisation of lectin protomers (i.e. lectin multivalency) (Lis and Sharon 1998).

The combination of lectin and ligand multivalency provides the basis for multiple high-affinity interactions that can lead to cross-linking (Rini and Lobsanov 1999). For example, crystallographic analysis of soybean agglutinin (SBA) with oligosaccharides revealed the existence of cross-linked lattices involving one SBA

tetramer and four pentasaccharides (Olsen et al. 1997). In addition, galectin-1 has been found to form chains of dimers that are cross-linked through oligosaccharides (Bourne et al. 1994; Perillo et al. 1995). Cross-linking is thought to be an important process and has been implicated in a variety of biological activities, for example apoptosis (Perillo et al. 1995).

### 3.1.2 Classification of Lectins

#### 3.1.2.1 Monosaccharide-Specificity

Lectins can be divided into five classes on the basis of their carbohydrate specificity according to the monosaccharide for which they exhibit the greatest affinity. All known lectins are specific for one of the following monosaccharides: GlcNAc; mannose (Man); fucose (Fuc); galactose/N-acetylgalactosamine (Gal/GalNAc); or N-acetylneuraminic acid (sialic acid). These monosaccharides are typically expressed on the surfaces of eukaryotic cells (Lis and Sharon 1998). Despite the binding affinity of lectins for monosaccharides being relatively low (i.e. in the millimolar range), some lectins are highly selective for the monosaccharide they bind (Lis and Sharon 1998). Most lectins also bind to di-, tri- and oligosaccharides, often with higher affinity than that with which they bind monosaccharides, and some lectins (for example potato lectin, selectins and galectins) only bind oligosaccharides (Lis and Sharon 1998). Lectins of the same specificity group often exhibit different binding affinities for oligosaccharides (Goldstein and Hayes 1976). For example, the siglecs all bind sialic acid, recognising a free hydroxyl group at position C9, but differ in their affinities for sialic acids modified at other positions (Kelm et al. 1998). Some lectins specific for the same oligosaccharide recognise different regions of its structure.

#### 3.1.2.2 Structural Differentiation

Alternatively, lectins may be grouped into three structural classes – simple lectins, mosaic (multi-domain) lectins or macromolecular assemblies. Simple lectins, such as the galectins and most plant lectins, are made up of a small number of subunits (protomers) and may have an additional domain besides the combining site. Mosaic lectins, for example influenza virus haemagglutinin and mannose-6-phosphate receptor, are membrane-bound proteins that consist of several kinds of protein domain, of which only one has a combining site. Macromolecular assemblies are filamentous organelles (fimbriae or pili) expressed on the surface of a bacterium. They consist of helically arranged subunits assembled in a well-defined order and only one of the subunits possesses a combining site. Within these three structural classes, lectins can

be further sub-divided into several families with similar sequences and structural properties (Lis and Sharon 1998). Examples of lectins from each structural family are presented in Table 2.

**Table 2 - Classification of Lectins According to Structure**

| Class | Family | Example |
|---|---|---|
| Simple lectins | Legume | Concanavalin A |
| | Cereal | Wheat germ agglutinin |
| | *Amaryllidaceae* | Snowdrop lectin |
| | *Moraceae* | Jacalin |
| | *Euphorbiaceae* | Ricin |
| | Galectins | Human galectin-7 |
| | Pentraxins | Human serum amyloid P component |
| Mosaic lectins | Viral haemagglutinins | Influenza virus haemagglutinin |
| | C-type | E-selectin |
| | P-type | Mannose-6-phosphate receptors |
| | I-type | Sialoadhesins |
| Macromolecular assemblies | | *E.coli* type 1 fimbriae |

All lectins can be classified into three groups on the basis of their structure – simple lectins, mosaic lectins and macromolecular assemblies. Within these three structural classes, lectins can be further sub-divided into families with similar sequences and structural properties. An example of each lectin family is provided.

### 3.1.3 Lectins Mediate a Variety of Biological Processes

Protein-carbohydrate interactions are involved in a large number of biological processes, predominantly involving cell-cell interactions. Glycoproteins are widespread in nature and their recognition is key to many cellular activities. Not only do lectins play roles in targeting glycoproteins to cellular locations, they are also implicated in immune responses and regulation of levels of proteins in circulation (Lee and Lee 1995).

#### 3.1.3.1 Animal Lectins

There are several classes of C-type lectin, the carbohydrate-binding properties of which are calcium-dependent. The collectins, such as mannan binding lectin (Ikeda et al. 1987) and ficolins are involved in innate immune responses by binding oligomannosides on infectious organisms and causing activation of the complement system and lysis of the pathogen (Holmskov et al. 2003). The mammalian hepatic lectins are receptors that mediate the internalisation and lysosomal degradation of galactose/GalNAc-containing glycoproteins (Ashwell and Harford 1982; Weigel 1994). Selectins are membrane proteins that mediate the adhesion of circulating leukocytes to

endothelial cells of blood vessels and the subsequent migration of these leukocytes into tissues (Lasky 1991). Endothelial leukocyte adhesion molecule 1, for example, mediates the adhesion of blood neutrophils (Bevilacqua et al. 1989). Leukocyte trafficking to sites of infection is an important regulatory process but can also cause harmful inflammation, for example in rheumatoid arthritis (Lis and Sharon 1998).

The I-type lectins play roles in cell adhesion and signalling. Sialoadhesin, for example, is active in the immune system (van den Berg et al. 1992). The P-type lectins are mannose-6-phosphate receptors that target lysosomal enzymes to the lysosome (reviewed in Hille-Rehfeld 1995). Pentraxins, including C-reactive protein and serum amyloid P component, regulate innate immunity against microbes and scavenge cellular debris, recognising and clearing apoptotic cells from circulation (reviewed in Gewurz et al. 1995).

The galectins (S-type animal lectins) are believed to function in cell adhesion and regulation of growth. As their expression is developmentally regulated, galectins have also been postulated to be essential for development and cell differentiation (Lis and Sharon 1998). Galectin-1 induces apoptosis in developing T cells in the thymus and activated T cells in the periphery (Perillo et al. 1995; Perillo et al. 1997; Vespa et al. 1999) and galectin-3 promotes the attachment of neutrophils to fibronectin and laminin (Kuwabara and Liu 1996).

### 3.1.3.2 Plant Lectins

Plant lectins have been known for longer than animal and microbial lectins and their carbohydrate-binding specificities and structures have been extensively studied. However, their biological functions remain the subject of much debate. The two most popular theories are that plant lectins are either involved in the specificity of legume root nodulation by bacteria or that they play roles in plant defence.

The lectin recognition hypothesis, proposed in the early 1970s (Bohlool and Schmidt 1974; Dazzo and Hubbell 1975; Hamblin and Kent 1973), suggests that lectins mediate specificity in the symbiotic relationship between nitrogen-fixing bacteria of the genus *Rhizobium* and leguminous plants. Indeed, the association between bacteria and legumes is highly specific, with the rhizobia exhibiting a very narrow host range. For example, rhizobia that infect and nodulate soybeans cannot nodulate peas or clover (Lis and Sharon 1998; van Rhijn et al. 1998) and it is thought that legume lectins facilitate the attachment of rhizobial bacteria to legume root hairs (Hirsch et al. 2001). However, whilst there is a great deal of support for this hypothesis, it does not explain the presence or role of lectins in non-leguminous plants. The other potential role of

lectins in plants is in defence against predators and pathogenic fungi (reviewed in Peumans and van Damme 1995). Animal lectins have been implicated in innate immunity and it is plausible that plant lectins might play a similar role. Plant lectins are known to be toxic to animals and inhibit fungal growth, which is the basis for this potential role in defence (Lis and Sharon 1998).

### 3.1.3.3 Bacterial and Viral Lectins

Lectins expressed on the surfaces of viruses and bacteria mediate adhesion of the organism to host cells as a first step in the infection process. The first example of this was the influenza virus haemagglutinin, which binds to gangliosides on the host cell surface to gain entry to the cell and replicate (Bergelson et al. 1982). Influenza virus haemagglutinin specifically binds to N-acetylneuraminic acid residues expressed on host cells, enabling the viral and cellular membranes to fuse as a prerequisite for viral infection. Adherence of the influenza haemagglutinin to neutrophils may allow the virus to spread from respiratory epithelial cells to lymph nodes (Ratcliffe et al. 1992). Structural knowledge of this lectin has enabled the rational design of an inhibitor that targets acts as a substrate analogue and can be used to treat influenza (Smith et al. 2001; von Itzstein et al. 1996).

Following the attachment of a pathogenic microbe to the host cell, it is thought that biological cross-talk via the lectin activates the host cell, leading to disease. For example, the attachment of *Escherichia coli* to intestinal epithelial cells leads to the transmigration of polymorphonuclear leukocytes to the infected tissue (Savkovic et al. 1996). In addition, secretion of chloride ions from epithelial cells is induced by the action of enterotoxins, produced by *E.coli*, that activate the cyclic GMP pathway and this causes diarrhoea (Uzzau and Fasano 2000).

### 3.1.3.4 Therapeutic Applications of Lectins

The specificity of lectins for glycoproteins can be exploited in medicine and research to distinguish between different cell types. For example, agglutination of B and T splenocytes by SBA is employed to purge human bone marrow for transplantation into children with severe combined immune deficiency or leukaemia (Aversa et al. 1994). The haemagglutinating activities of lectins are also used in blood-typing. Furthermore, mitogenic lectins (for example phytohaemagglutinin) that activate lymphocytes and cause them to divide are used to assess immunocompetence. Several lectins can be used in combination to stain urinary cell types to monitor kidney function post-transplantation (Grupp et al. 2002; Grupp et al. 2001). P-fimbriated *E.coli*

that recognise and bind to uroepithelial cells can be used as a marker to detect risk groups among patients with recurrent urinary tract infections (Jacobson et al. 1986).

Animal and plant lectins are also thought to be of potential use in targeted drug delivery, particularly for the treatment of cancer and immune diseases (Clark et al. 2000; Lehr 2000). Selectins are thought to be of potential use in targeting anti-inflammatory drugs to sites of inflammation (Eniola et al. 2002). Wheat germ agglutinin (WGA) has potential as a vehicle for the delivery of modified liposomes in oral vaccines (Chen et al. 1996). It binds a variety of cell types (Aubery et al. 1990; Gabor et al. 1998; Lotan et al. 1975) and has been used to target chemotherapeutic prodrugs to colon carcinoma cells (Wirth et al. 1998).

### 3.1.4 Legume Lectins

The legume lectins are mainly isolated from the seeds of leguminous plants (Loris et al. 1998). Members of this protein family are highly homologous with respect to their amino acid sequences and share a common tertiary structure but exhibit structural variation at the quarternary level.

Legume lectins are typically synthesised as precursor polypeptides, which are post-translationally cleaved, beginning with the removal of an N-terminal signal peptide, to yield the mature protein (Loris et al. 1998). Legume lectin protomers have molecular weights in the range 25-30kDa and comprise a polypeptide chain of ~250 amino acids. Each protomer contains two metal ions, calcium and manganese, located close to the combining site that are required for carbohydrate binding. All known legume lectins are active as either dimers or tetramers, although each protomer possesses a functional carbohydrate-combining site with the combining site in each protomer having the same specificity (Sharon and Lis 2002). Many legume lectins are glycoproteins, in which the protomers carry one or two N-linked oligosaccharides (Lis and Sharon 1998). The presence of bound oligosaccharides has been postulated to influence the quarternary structures of the legume lectins (discussed later).

#### 3.1.4.1 Sequence Homology Among Legume Lectins

There is high sequence homology among the legume lectins, with members of this protein family sharing up to 99% sequence identity (Manoj and Suguna 2001). An alignment of the sequences of ten legume lectins is presented in Figure 10. Analysis of 19 legume lectins of known tertiary structure revealed that 69 amino acids were conserved, of which 21 were identical, among their sequences (Manoj and Suguna 2001). The residues conserved in all legume lectin sequences are involved in carbohydrate binding and coordination of the calcium and manganese ions that are

bound close to the combining site (Lis and Sharon 1998). It is interesting to note that there are ten amino acids different between the sequence of *rec*ECL determined for this investigation by DNA sequencing (Stancombe et al 2003) and that reported for the *Sven*ECL (Svensson et al 2002).

**Figure 10 - Sequence Alignment of Ten Legume Lectins**

```
PSL      ---MASLQTQ MISFYAIFLS ILLTTILFFK VNSTETTSFL ITKFS----P DQQNLIFQGD
LenL     ---------- ---------- ---------- ---TETTSFS ITKFS----P DQQNLIFQGD
SBA      -MATSKLKTQ NVVVSLSLTL TLVLVLLTSK ANSAETVSFS WNKFV----P KQPNMILQGD
GSIV     ---------- ---------- ---------- ---QNTVNFT YPDFWSYSLK NGTEITFLGD
PHA-L    ---------- ---MASSKFF TVLFLVLLTH ANSSNDIYFN FQRFN----- -ETNLILQRD
ECorL    -------MAT YKLCSVLALS LTLFLLILNK VNSVETISFS FSEFE----P GNDNLTLQGA
ECL      ---------- ---------- ---------- ---VETISFS FSEFE----P GNNDLTLQGA
SvenECL  ---------- ---------- ---------- ---VETISFS FSEFE----P GNDNLTLQGAWBA-1
---------- ---------- ---------- ---MKTISFN FNQFH----Q NEEQLKLQRD
ConA     MAISKKSSLF LPIFTFITMF LMVVNKVSSS THETNALHFM FNQFS----K DQKDLILQGD
PNA      ---------M KPFCVFLTFF LLLA-ASSKK VDSAETVSFN FNSFS----E GNPAINFQGD

PSL      GYTTKEK-LT LTKAVKN--- ---TVGRALY SSPIHIWDRE TGNVANFVTS FTFVIN-APN
LenL     GYTGKEG-LT LTKVSKE--- ---TGGRALY STPIHIWDRD TVNVANFVTN GSQVFRESPN
SBA      AIVTSSGKLQ LNKVDENGTP KPSSLGRALY STPIHIWDKE TGSVASFAAS FNFTFY-APD
GSIV     ATR-IPGALQ LTKTDANGNP VRSSAGQASY SEPVFLWD-S TGKAASFYTS FTFLLK--NY
PHA-L    ASVSSSGQLR LTNLNGNGEP RVGSLGRAFY SAPIQIWDNT TGTVASFATS FTFNIQ-VPN
ECorL    ALITQSGVLQ LTKINQNGMP AWDSTGRTLY AKPVHIWDMT TGTVASFETR FSFSIEQPYT
ECL      AIITQSGVLQ LTKINQNGMP AWDSTGRTLY TKPVHIWDMT TGTVASFETR FSFSIEQPYT
SvenECL  ALITQSGVLQ LTKINQNGMP AWDSTGRTLY TKPVHMWDST TGTVASFETR FSFSIEQPYTWBA-1
ARISSNSVLE LTKV-VNGVP TWNSTGRALY AKPVQVWDST TGNVASFETR FSFSIRQPFP
ConA     ATTGTEGNLR LTRVSSNGSP QGSSVGRALF YAPVHIWESS -AVVASFEAT FTFLIK-SP-
PNA      VTVLSNGNIQ LTNLN----- KVNSVGRVLY AMPVRIWSSA TGNVASFLTS FSFEMK-DIK

PSL      SYNVADGFTF FIAPVDTKPQ TGGGYLG--- ---------- -VFNS-AEYD KTT-QTVAVE
LenL     GYNVADGFTF FIAPVDTKPQ TGGGYLG--- ---------- -VFYNGKEYD KTS-QTVAVE
SBA      TKRLADGLAF FLAPIDTKPQ THAGYLG--- ---------- -LFNE---NE SGD-QVVAVE
GSIV     GAPTADGLAF FLAPVDSSVK DYGGFLG--- ---------- -LFRHETAAD PSKNQVVAVE
PHA-L    NAGPADGLAF ALVPVGSQPK DKGGFLG--- ---------- -LFDG----S NSNFHTVAVE
ECorL    RPLPADGLVF FMGPTKSKPA QGYGYLG--- ---------- -IFNN--SKQ DNSYQTLGVE
ECL      RPLPADGLVF FMGPTKSKPA QGYGYLG--- ---------- -VFNN--SKQ DNSYQTLAVE
SvenECL  RPLPADGLVF FMGPTKSKPA QGYGYLG--- ---------- -VFNN--SKQ DNSYQTLAVE
WBA-1    RPHPADGLVF FIAPPNTQTG EGGGYFG--- ---------- -IYN-----P LSPYPFVAVE
ConA     DSHPADGIAF FISNIDSSIP S--GSTGRLL GLFPDANVIR NSTTIDFNAA YNADTIVAVE
PNA      DYDPADGIIF FIAPEDTQIP A--GSIG--- ---------- -GGTLGVSDT KGAGHFVGVE

PSL      FDTFYNAAWD PSNRDRHIGI DVNSIKSVNT KSWKLQNGE- --EANVVIAF NAATNVLTVS
LenL     FDTFYNAAWD PSNKERHIGI DVNSIKSVNT KSWNLQNG-- ---------- ----------
SBA      FDTFRN-SWD PPNP--HIGI NVNSIRSIKT TSWDLANNK- --VAKVLITY DASTSLLVAS
GSIV     FDTWINKDWN DPPYP-HIGI DVNSIVSVAT TRWENDDAYG SSIATAHITY DARSKILTVL
PHA-L    FDTLYNKDWD PTER--HIGI DVNSIRSIKT TRWDFVNGE- --NAEVLITY DSSTNLLVAS
ECorL    FDTFSN-PWD PPQVP-HIGI DVNSIRSIKT QPFQLDNGQ- --VANVVIKY DASSKILHAV
ECL      FDTFSN-PWD PPQVP-HIGI DVNSIRSIKT QPFQLDNGQ- --VANVVIKY DASSKILLAV
SvenECL  FDTFSN-PWD PPQVP-HIGI DVNSIRSIKT QPFQLDNGQ- --VANVVIKY DAPSKILHVV
WBA-1    FDTFRN-TWD P-QIP-HIGI DVNSVISTKT VPFTLDNGG- --IANVVIKY DASTKILHVV
ConA     LDTYPNTDIG DPSYP-HIGI DIKSVRSKKT AKWNMQNGK- --VGTAHIIY NSVDKRLSAV
PNA      FDTYSNSEYN DPPTD-HVGI DVNSVDSVKT VPWNSVSGA- --VVKVTVIY DSSTKTLSVA
```

```
PSL      LTYPNSLEEE NVTSYTLSDV VSLKDVVPEW VRIGFSATTG ----AEY--- AAHEVLSWSF
LenL     ---------- -VTSYTLNEV VPLKDVVPEW VRIGFSATTG ----AEF--- AAQEVHSWSF
SBA      LVYPS----- QRTSNILSDV VDLKTSLPEW VRIGFSAATG ----LDIP-G ESHDVLSWSF
GSIV     LSYEHG---- --RDYILSHV VDLAKVLPQK VRIGFSAGVG ----YDE--- -VTYILSWHF
PHA-L    LVYPSQ---- -KTSFIVSDT VDLKSVLPEW VSVGFSATTG ----INKGNV ETNDVLSWSF
ECorL    LVYPSS---- -GAIYTIAEI VDVKQVLPEW VDVGLSGATG ----AQRDAA ETHDVYSWSF
ECL      LVYPSS---- -GAIYTIAEI VDVKQVLPEW VDVGLSGATG ----AQRDAA ETHDVYSWSF
SvenECL  LVYPSS---- -GAIYTIAEI VDVKQVLPDW VDVGLSGATG ----AQRDAA ETHDVYSWSF
WBA-1    LVFPSL---- -GTIYTIADI VDLKQVLPES VNVGFSAATG DPSGKQRNAT ETHDILSWSF
ConA     VSYPNA---- --DSATVSYD VDLDNVLPEW VRVGLSASTG ------LY-K ETNTILSWSF
PNA      VTNDNG---- --DITTIAQV VDLKAKLPER VKFGFSASGS ------LGGR QIHLIRSWSF

PSL      HSELSGTSSS KQAADA---- ------
LenL     NSQLGHTSKS ---------- ------
SBA      ASNLPHASSN IDPLDLTSFV LHEAI-
GSIV     FSTLDGTNK- ---------- ------
PHA-L    ASKLSDGTTS EG-LNLANLV LNKIL-
ECorL    QASLPETNDA VIPTSNHNTF AI----
ECL      HASLPETN-- ---------- ------
SvenECL  QASLPE---- ---------- ------
WBA-1    SASLPGTNEF ---------- ------
ConA     TSKLKS---- ----NE--IP DIATVV
PNA      TSTLITTTRR SIDNNEKKIM NMASA-
```

Alignment of the amino acid sequences of ten legume lectins whose three-dimensional structures have been determined was performed using clustalw (Thompson et al. 1994) with sequences taken from the SwissProt database (Boeckmann et al. 2003). The sequence of ECL quoted above is that of *rec*ECL, determined for this investigation by DNA sequencing (Stancombe et al 2003), whilst the *Sven*ECL sequence is that of the recently reported crystal structure of ECL (Svensson et al 2002). Six sequences include the N-terminal signal sequence that is post-translationally cleaved to yield the mature protein. Legume lectins are highly homologous in terms of their primary structures, sharing up to 99% sequence identity. In this figure, the amino acid residues highlighted in blue are identical among the sequences and those highlighted in yellow are conservatively substituted. The conserved residues are involved in coordinating the calcium and manganese ions as well as in binding carbohydrates.

(Abbreviations: PSL, pea lectin; LenL, lentil lectin; SBA, soybean agglutinin; GSIV, *Griffonia simplicifolia* lectin 4; PHA-L, phytohaemagglutinin; ECorL, *Erythrina corallodendron* lectin; ECL, *Erythrina cristagalli* lectin; WBA-1 winged bean acidic lectin; ConA, concanavalin A; and PNA, peanut agglutinin).

---

Carbohydrate binding requires three amino acids – an aspartic acid, an asparagine and an aromatic residue or leucine – that are invariant among all the legume lectins, regardless of their monosaccharide specificity (Lis and Sharon 1998). The Asp and Asn residues are also involved in coordinating the bound calcium ion, which explains the requirement of this metal ion for carbohydrate binding by legume lectins. Differences in carbohydrate specificity are conferred by variation in the amino acid residues, other than the conserved triad, that line the shallow cleft that makes up the combining site.

### 3.1.4.2 The Legume Lectin Fold

**Figure 11 - The Jelly-roll Motif (Legume Lectin Fold)**


This figure, prepared using the programs Molscript (Kraulis 1991), Povray (www.povray.org/) and GIMP (www.gimp.org/), shows the three-dimensional structure of Concanavalin A (Reeke et al. 1975; PDB code 2CNA). The overall topology is described as a jelly-roll and is also known as the legume lectin fold. The structure comprises three β-sheets: one flat, six-stranded anti-parallel β-sheet at the back of the lectin in this figure; a curved, seven-stranded anti-parallel β-sheet at the front of the molecule; and a small, five-stranded β-sheet at the bottom. These sheets are joined by a series of loops. A typical legume lectin, Con A contains two metal ions (calcium and manganese, shown in green and navy, respectively) that are required for its carbohydrate-binding activity. The metal ions are bound close to the combining site, which is located in a shallow depression on the surface of the curved, seven-stranded β-sheet.

The first lectin to be structurally characterised was concanavalin A (ConA), which is isolated from the jack bean, *Canavalia ensiformis*, a leguminous plant (Hardman and Ainsworth 1972). The overall fold, characterised by two anti-parallel β-sheets, is described as the "jelly-roll" motif (Figure 11). This toplogy has since been

observed in many lectins and is also referred to as the legume-lectin fold. The legume lectin fold comprises three β-sheets connected by a series of loops. As depicted in Figure 11, towards the back of the structure is a flat, six-stranded sheet and towards the front is a curved, seven-stranded sheet, both of which are entirely anti-parallel. There is also a small five-membered β-sheet, seen at the bottom of the molecule in Figure 11. The carbohydrate-combining site is located in a shallow depression on the surface of the lectin and involves four loops associated with the curved, seven-stranded β-sheet. In each protomer, there are two metal ions, $Ca^{2+}$ and $Mn^{2+}$, bound close to the legume lectin combining site that are essential for the carbohydrate binding activity of these proteins.

The $Ca^{2+}$ and $Mn^{2+}$ are thought to maintain the correct spatial arrangement of combining site residues (Derewenda et al. 1989). The metal ions are each coordinated by the side chains of four amino acids and two water molecules. The manganese ion in SBA can be substituted with other transition metals (nickel, cobalt and cadmium) with no loss of carbohydrate-binding activity (Jaffe et al. 1977). Furthermore, the crystal structures of ConA in which the manganese ion has been replaced by cobalt or cadmium reveal that the overall geometry of the metal binding site is independent of the nature of the transition metal present (Emmerich et al. 1994).

In legume lectins, the $Ca^{2+}$ and $Mn^{2+}$ ions also stabilise a conserved *cis*-peptide bond between the carbohydrate-binding aspartate residue and the preceding amino acid, which is typically alanine. The stereochemistry of the *cis*-peptide bond places the side chain of the aspartate residue in the correct orientation for carbohydrate binding (Lis and Sharon 1998). Interestingly, the structures of metal-stripped ConA determined at different stages of metal replenishment have revealed a *trans* to *cis* isomerisation of the Ala-Asp peptide bond upon metal-induced activation of the lectin (Bouckaert et al. 1995; Bouckaert et al. 1996).

Legume lectins are classed as simple lectins and, although they are all biologically active as dimers or tetramers, the association of protomers does not affect carbohydrate-binding because each protomer is equipped with a fully-functional, pre-formed combining site. However, it is interesting that the legume lectins exhibit variation in their quarternary association considering the high-level of homology among their structures at the primary and tertiary levels.

### 3.1.4.3 Variation in Quarternary Association Among Legume Lectins

Despite a high degree of similarity among their tertiary structures the legume lectins exhibit a range of quarternary structures. All legume lectins are biologically

active as dimers or tetramers (which are considered dimers of dimers). There are five types of interface, all of which involve the flat, six-stranded β-sheet. In II-type interfaces, protomers associate side-by-side and in X1, X2, X3 and X4 type interfaces, protomers associate back-to-back.

The canonical mode of dimerisation was originally observed for ConA and involves the side-by-side (II-type) association of two protomers such that the two six-stranded β-sheets form one contiguous 12-stranded β-sheet (Hardman and Ainsworth 1972). The X3-type of dimer, first observed in *Erythrina corallodendron* lectin (ECorL), is described as the "handshake motif" because protomers associate back-to-back with the six-stranded sheet from one protomer tilted with respect to the other (Shaanan et al. 1991). Protomers of the lectin isolated from *Griffonia simplicifolia* (GSIV) also dimerise back-to-back, but in a different way to ECorL, forming an X4-type interface (Delbaere et al. 1993).

X1 and X2-type interfaces are observed in tetrameric lectin structures. In the lectin isolated from *Dolichos biflorus*, two X1-type dimers associate side-by-side, forming a II-type interface (Hamelryck et al. 1999). ConA can exist as a dimer and as a tetramer, with the tetrameric structure formed by two II-type dimers associating back-to-back, forming an X2-type interface (Bouckaert et al. 1995; Reeke et al. 1975). Peanut lectin (PNA) has the most unusual quarternary structure, described as an "open" arrangement. Unlike in the other tetrameric lectins, the protomers of peanut lectin are arranged as two back-to-back dimers with interactions between the dimers involving different interfaces (Banerjee et al. 1994). Although, the two dimers associate side-by-side, they do not form a contiguous 12-stranded β-sheet but are connected by a set of water bridges.

Originally, the observed variation in quarternary association was explained by the presence of N-linked glycosylation close to residues that would be involved in the formation of a canonical dimer interface and to small differences in primary structure. The presence of a branched oligosaccharide N-linked to $Asn^{17}$ of ECorL was deemed responsible for forcing the protomers into a non-canonical dimer structure (Shaanan et al. 1991). However, the glycosylation sites in basic winged bean lectin, which shares 63% sequence identity with ECorL and forms X3-type dimers (Prabu et al. 1998), are distant from the region that would constitute the interface in a canonical dimer. In addition, peanut lectin is not glycosylated but the dimeric interface of the tetramer is non-canonical. Taken together, these data suggest that factors other than the presence of glycosylation influence the mode of quarternary association in legume lectins (Prabu et al. 1999).

Assessment of the different interfaces in terms of hydrophobic surface area buried upon association, interaction and energy, and shape complementarity revealed that the native mode of dimerisation or tetramerisation for each of the lectins studied was the most favourable arrangement (Prabu et al. 1999). Analysis of the primary structures of several legume lectins has since pinpointed a small number of residues that appear to play a role in determining certain modes of dimerisation (Manoj and Suguna 2001). However, it is not yet possible to accurately predict the quarternary structures of uncharacterised lectins.

### 3.1.4.4 Carbohydrate Binding in Legume Lectins

The architecture of the combining site is conserved among legume lectins, as are many of the amino acids involved in carbohydrate binding (Sharma and Surolia 1997). The combining site comprises a shallow cleft involving four loops (designated A-D) that are associated with the curved, seven-stranded β-sheet. Two conserved residues, an asparagine and an aspartic acid, are essential for carbohydrate binding along with the main chain nitrogen of a glycine or arginine (Sharma and Surolia 1997). These residues form hydrogen bonds with the monosaccharide in the combining site. Sugar binding is further stabilised by stacking interactions with either a leucine or a hydrophobic residue. The conserved aspartic acid and glycine/arginine are present on combining site loops A and B, respectively (Sharma and Surolia 1997). The asparagine and leucine/hydrophobic residue are located in loop C and loop D provides residues that interact with the bound monosaccharide via the peptide backbone. Analysis of the sequences of 26 legume lectins revealed a link between the length of combining site loop D and monosaccharide specificity, whereby the length of the loop alters the shape of the combining site (Sharma and Surolia 1997). For example, legume lectins that are specific for galactose/GalNAc have longer D loops than those that recognise mannose, glucose or fucose.

Legume lectins are highly selective for the monosaccharides they bind. For example, those that are galactose-specific do not bind Man or glucose, even though they are epimers of Gal. Monosaccharides can only bind in an orientation that directs an equatorial hydroxyl group towards the space between the Asp-Gly/Arg-Asn triad and provides an axial hydroxyl or hydroxymethyl group as a hydrogen-bonding partner for the aspartic acid (Sharma and Surolia 1997). Lectins bind carbohydrates by forming a network of hydrogen bonds and hydrophobic interactions (Lis and Sharon 1998). The hydroxyl groups on the sugar form hydrophobic patches that interact with hydrophobic regions on the lectin, for example the stacking interactions observed between pyranose

rings and aromatic side chains. It has also been shown that water molecules bridge hydrogen bonds between carbohydrate and lectin. For example, one of the waters that co-ordinates $Ca^{2+}$ also forms a bridge with the carbonyl group of the carbohydrate-binding aspartic acid (Adar et al. 1998; Loris et al. 1994).

### 3.1.5 *Erythrina Cristagalli* Lectin

The lectin isolated from seeds of the coral tree *Erythrina cristagalli* is known as ECL or ECA. Trees and shrubs of the genus *Erythrina* are deciduous legumes found in tropical and subtropical habitats (Lis and Sharon 1987). Lectins isolated from *Erythrina* species are glycoproteins, comprising 3-10% carbohydrate, with molecular weights in the range 56-68kDa. *Erythrina* lectins are specific for Gal- or GalNAc-containing carbohydrates and are biologically active as dimers (Lis and Sharon 1987). They are typical legume lectins, with protomers that adopt the legume lectin fold and contain one calcium and one manganese ion. All *Erythrina* lectins agglutinate human erythrocytes of all blood groups and most are mitogenic for human peripheral blood lymphocytes (Lis and Sharon 1987).

#### 3.1.5.1 Carbohydrate Specificity of ECL

ECL selectively binds to terminal unsubstituted N-acetyllactosamine (LacNAc) (Iglesias et al. 1982). *In vitro*, 0.4mM LacNAc completely inhibits the activity of four agglutinating units of ECL (one unit of activity is defined as the lowest concentration of lectin giving visible haemagglutination). ECL binds lactose, GalNAc and Gal with lower affinity than LacNAc, with the binding affinity for these sugars decreasing in the order they are listed (Iglesias et al. 1982). However, it has recently been shown that ECL has the highest affinity for fucosyllactose and fucosyllactosamine, binding both sugars with similar affinity. In this respect, ECL differs from lectins isolated from other *Erythrina* species, which exhibit a preference for fucosyllactose (Moreno et al. 1997).

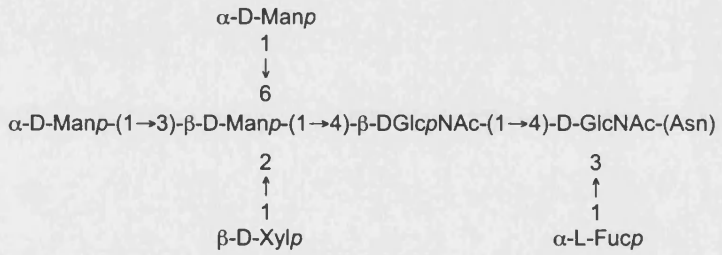
The biological function of ECL in the legume is currently unknown. *In vitro*, ECL agglutinates human erythrocytes of all blood types and is mitogenic for human peripheral blood T lymphocytes (Iglesias et al. 1982). ECL might be a useful marker to follow changes in cell surface carbohydrates during differentiation. For example, ECL recognises oligosaccharides expressed on human cervical carcinoma cells that are not expressed on healthy cervical cells (Banerjee et al. 1995). More recently, ECL has been used to select subpopulations of insect cells by the character of their cell-surface oligosaccharides (Donaldson and Shuler 1999).

### 3.1.5.2 Molecular Structure of ECL

ECL is expressed as a precursor protein that is post-translationally processed (involving the removal of the 26-amino acid signal sequence from the N-terminus and addition of oligosaccharide) to yield the mature protein. The sequence of ECL is 96% identical to that of ECorL – there are only nine differences between their 241-amino acid sequences. Unlike ECorL, the ECL sequence contains a single potential N-linked glycosylation site at residue $Asn^{113}$ and the molecular weight of the unglycosylated protomer is predicted to be 26kDa (Stancombe et al. 2003).

ECL is biologically active as a dimer of ~56.8kDa. Sodium dodecylsulphate polyacrylamide gel electrophoresis (SDS-PAGE) of ECL produces two bands that both stain positively for protein and carbohydrate (Iglesias et al. 1982). This indicates that ECL is a heterodimer, made up of two protomers that are both glycosylated, with a single carbohydrate chain bound to each. The oligosaccharides bound to ECL have been characterised with the major component identified as a heptasaccharide of 8.8 glucose units (Figure 12) and minor components of 8.0, 7.2 and 5.8 glucose units (Ashford et al. 1991). The glycosylation profile of ECL matches those of four other lectins isolated from *Erythrina* species and five unrelated legume lectins.

**Figure 12 - Glycosylation Profile of ECL**




The major component of the oligosaccharides isolated from ECL is the heptasaccharide diagrammatically represented in this figure. This glycosylation profile has also been observed in four other *Erythrina* lectins and five unrelated legume lectins.

### 3.1.5.3 Crystal Structure of ECL

The crystal structures of ECL in complex with 2'-α-L-fucosyllactose and lactose (PDB codes 1GZ9 and 1GZC, respectively) were recently reported (Svensson et al. 2002). Four sequences for ECL have been reported (Svensson et al. 2002). The sequence determined by mass spectrometry contained only 229 out of 241 amino acids and differed at two positions from the sequence of *Sven*ECL according to the observed electron density (Svensson et al. 2002). Peptide mapping was reportedly

attempted by analysing fragments of *Sven*ECL (created by enzymatic digestion of the protein) using tandem mass spectrometry. However, sequence coverage of only 89% of the amino acid residues (i.e. 213 of 239) was obtained by this method (Svensson et al 2002).

The *Sven*ECL protomer was confirmed to adopt the jelly-roll fold and protomers were shown to dimerise back-to-back via the handshake motif. N-linked glycosylation was apparently observed at two positions in the structure, $Asn^{17}$ and $Asn^{113}$, although the sugars were not modelled into the observed electron density. The influence of bound carbohydrate on the quarternary structures of legume lectins is still the subject of much debate. In ECorL, which also forms dimers via the handshake motif, the heptasaccharide bound to $Asn^{17}$ was proposed to sterically hinder the formation of a canonical dimer and the same reasoning has been applied to *Sven*ECL (Svensson et al. 2002).

In the *Sven*ECL combining site, the side-chains of $Asp^{89}$, $Gly^{107}$ and $Asn^{133}$ create a strong hydrogen-bonding network for carbohydrate binding, a structural water molecule mediates interaction between the main chain carbonyl oxygen of $Leu^{86}$ and galactose and the side-chain of $Phe^{131}$ forms hydrophobic stacking interactions with the galactose pyranose ring. Residues $Ala^{218}$ and $Gln^{219}$ also form hydrogen bonds with the bound carbohydrate. On the basis of the crystallographic data, it was proposed that observed differences in the binding affinities of ECL and ECorL for the same carbohydrates are due to differences in the amino acid sequences of the two proteins at positions 111 and 125 (Svensson et al. 2002). It is thought that substitutions at these positions cause the side chain of $Val^{92}$ to rotate, transmitting structural changes to the combining site, although this remains to be proved.

## 3.2 MATERIALS AND METHODS

### 3.2.1 Materials

Native and recombinant forms of ECL (*n*ECL and *rec*ECL, respectively) were provided by Dr J.A. Chaddock (Health Protection Agency, UK). *n*ECL was purchased from Sigma (UK) and *rec*ECL was expressed and purified by the Toxin Therapeutics Division at CAMR as reported (Stancombe et al. 2003) and briefly described in Appendix 1. The amino acid sequence of *rec*ECL (Figure 13) does not differ from that of *n*ECL. Purified protein stocks were stored at -20°C in buffer comprising 50mM sodium chloride and 20mM HEPES, pH 7.4. The protein concentrations of the *n*ECL and *rec*ECL stocks were estimated as ~10.0mg/ml and ~14.5mg/ml, respectively.

**Figure 13 - Amino Acid Sequence of *rec*ECL compared with that of *Sven*ECL**

| | |
|---|---|
| ***n*ECL** | VETISFSFSE FEPGNNDLTL QGAAIITQSG VLQLTKINQN |
| ***Sven*ECL** | VETISFSFSE FEPGNDNLTL QGAALITQSG VLQLTKINQN |
| ***n*ECL** | GMPAWDSTGR TLYTKPVHIW DMTTGTVASF ETRFSFSIEQ |
| ***Sven*ECL** | GMPAWDSTGR TLYTKPVHMW DSTTGTVASF ETRFSFSIEQ |
| ***n*ECL** | PYTRPLPADG LVFFMGPTKS KPAQGYGYLG VFNNSKQDNS |
| ***Sven*ECL** | PYTRPLPADG LVFFMGPTKS KPAQGYGYLG VFNNSKQDNS |
| ***n*ECL** | YQTLAVEFDT FSNPWDPPQV PHIGIDVNSI RSIKTQPFQL |
| ***Sven*ECL** | YQTLAVEFDT FSNPWDPPQV PHIGIDVNSI RSIKTQPFQL |
| ***n*ECL** | DNGQVANVVI KYDASSKILL AVLVYPSSGA IYTIAEIVDV |
| ***Sven*ECL** | DNGQVANVVI KYDAPSKILH VVLVYPSSGA IYTIAEIVDV |
| ***n*ECL** | KQVLPEWVDV GLSGATGAQR DAAETHDVYS WSFHASLPET N |
| ***Sven*ECL** | KQVLPDWVDV GLSGATGAQR DAAETHDVYS WSFQASLPE- - |

The ECL gene devoid of the 26-amino acid N-terminal signal peptide was recombinantly over-expressed in *E.coli* (Stancombe et al, 2003) and the 241-amino acid sequence of the mature *n*ECL protein is presented in black type. The amino acid sequence reported for *Sven*ECL is presented in blue.

The sequence of ECL is highly homologous to the sequences of lectins isolated from other *Erythrina* species and is 96% identical to ECorL. Unlike ECorL, the ECL sequence contains only one potential site for N-linked glycosylation at residue $Asn^{113}$ (underlined). Analysis of the structures of ECL in complex with lactose and fucosyllactose (Svensson et al, 2002) has shown that residues $Leu^{86}$, $Asp^{89}$, $Gly^{107}$, $Phe^{131}$, $Asn^{133}$, $Ala^{218}$ and $Gln^{219}$ (highlighted in red in the *rec*ECL sequence) are involved in carbohydrate binding.

24-well crystallisation plates (Molecular Dimensions Ltd, UK) were used with siliconised coverslips (Hampton Research, USA). Commercial crystallisation screens Structure Screen I and II (Molecular Dimensions Ltd) and a PEG/Ion matrix (Hampton Research) were used in initial crystallisation trials. The contents of these screens are listed in Appendix 2. Reagents used to prepare the crystallisation solutions were of high quality and purity and solutions were made with AnalaR water (VWR International, UK). Sodium chloride and ethanol were purchased from Fisher Chemicals (UK) and

tert-butanol was obtained from Fluka (UK). Tris, isopropanol and glycerol were obtained from VWR International. All other reagents were purchased from Sigma.

Co-crystallisation experiments involved GalNAc; α-lactose (Lac); β-lactose; Gal; 2'fucosyllactose and 3'-fucosyllactose. A 100mM stock solution of each carbohydrate was prepared with milliQ water.

### 3.2.2 Crystallisation

Crystallisation was attempted using the hanging drop vapour diffusion method. Drops comprising 2μl of protein and 2μl of mother liquor were equilibrated against 800μl of mother liquor at 16°C. For co-crystallisation, 0.4μl of a 100mM carbohydrate solution was also added to the drop. Initial crystallisation screens were set up using *rec*ECL. Structure Screen I, Structure Screen II, and a polyethylene glycol (PEG)/Ion matrix were used to sample a wide range of potential crystallisation conditions. In addition, 2-methyl-2,4-pentanediol (MPD) was tested a precipitant, in a trial based on JBScreens 7 and 8 (JenaBioScience, Germany; Appendix 2). Promising conditions were repeated and optimised for both *n*ECL and *rec*ECL to produce crystals for X-ray diffraction experiments. Co-crystallisation with GalNAc, Lac and β-lactose was attempted using conditions reported for the crystallisation of ECorL (Shaanan et al. 1991). Further attempts were also made to co-crystallise both forms of ECL with carbohydrates.

### 3.2.3 Data Collection and Processing

X-ray diffraction data were collected at ~100K using the SRS at Daresbury Laboratory (UK) on stations PX14.1 (λ = 1.488Å) and PX14.2 (λ = 0.978Å). X-rays for both stations are produced using a multipole wiggler, focussed with a rhodium-coated mirror and filtered using a silicon monochromator. Diffraction images were recorded on ADSC Quantum 4R CCD area detectors.

Each crystal was soaked in cryoprotectant solution (comprising mother liquor with 20-25% glycerol) for a few minutes before being mounted in a wire loop. The loop was then placed on a goniometer head and the crystal positioned in the centre of the X-ray beam under a cold stream of nitrogen gas. A few test images were recorded for each crystal to determine the symmetry of the crystal lattice, the maximum resolution and the data collection strategy to be used.

A total of six ECL crystal forms were analysed and all diffracted to high resolution. Diffraction images collected from crystals of *rec*ECL (space group P1), *rec*-Lac (C2), *rec*-Lac ($P2_1$) and *rec*ECL ($P2_1$) were processed, scaled and truncated at

Daresbury Laboratory using the program suite HKL2000 (Otwinowski and Minor 1997). Diffraction images for the two remaining crystals were processed and scaled in house using the programs DENZO and SCALEPACK (Otwinowski and Minor 1997) with intensities truncated to amplitudes by the program MTZ2XPLOR of the CCP4 suite (Bailey 1994).

### 3.2.4 Molecular Replacement

Molecular replacement was the method chosen to determine initial phases for the structure of ECL. ECorL was used as a search model. ECorL is homologous to ECL – the two lectins share 96% sequence identity – and its crystal structure has been determined (Elgavish and Shaanan 1998; PDB code 1AX0).

For each dataset Vm was calculated to estimate the solvent content and number of protein molecules per asymmetric unit. Molecular replacement was then performed using the program AMoRe (Navaza 1994), which performs rotation and translation searches, rigid-body refinement and generates structure factors from a model in a large P1 cell. Solutions for the cross-rotation and translation functions were ranked according to the calculated correlation coefficient (CCF) and R factor (RF) of each peak. As a general rule, the top peak (that with the highest CCF and lowest RF) from each search was fixed in the subsequent translation search.

#### 3.2.4.1 *n*ECL

Initial phase angles were taken from the crystal structure of ECorL. The model was truncated to comprise only the coordinates of amino acids 3-239 from one ECorL protomer (i.e. all heteroatoms, ligands and water molecules were removed). The side chains of seven of the nine non-conserved residues ($Asp^{16}$, $Asn^{17}$, $Leu^{25}$, $Ile^{111}$, $Gln^{134}$, $His^{180}$, $Gln^{234}$) in the ECorL model were mutated to alanine using 6d_moleman from the CCP4 suite of programs (Bailey 1994). It was not necessary to alter the side chains of residues $Ala^{54}$ and $Gly^{125}$ in the ECorL model because they are small and unlikely to affect the success of the molecular replacement method.

The resolution range for the rotation search was limited to 10-3.0Å. The cell model had dimensions 78 × 63 × 58Å and the sphere of integration was of 15Å radius.

The molecular replacement solutions were viewed in O (Jones et al. 1991) to ensure that there were no steric clashes between the molecules or symmetry-related molecules and the coordinates of the two molecules were then written out as PDB files. The separate chains were relabelled (chain A and chain B) and consolidated into a single PDB file that was used as the initial model for refinement.

#### 3.2.4.2 *rec*ECL Structures

The refined coordinates of one *n*ECL protomer (with all heteroatoms, ligands and waters removed) were used as a model for molecular replacement for the *rec*ECL structures. The resolution range used for the rotation search was limited to 10-3.0Å, the cell model was 85 × 69 × 65Å and the sphere of integration had a radius of 15Å.

### 3.2.5 Refinement and Model Building

The PDB file generated for each crystal structure after molecular replacement was refined using the Crystallography & NMR Suite of programs (Brunger et al. 1998; http://cns.csb.yale.edu/v1.0/) to produce an initial structural model and electron density maps for analysis. Coordinate and structure factor files were generated for the initial model and a "test-set" (comprising ~5% of the reflections) was created to calculate the $R_{free}$ value for cross-validation (Brunger 1992). Two rounds of rigid body refinement were performed (Adams et al. 1997; Pannu and Read 1996): the first with all atoms refined as one rigid body and the second with each protomer refined as a separate rigid body. This was followed by crystallographic conjugate gradient minimisation refinement; restrained, individual B-factor refinement; and crystallographic simulated annealing refinement at 1500K (Adams et al. 1997; Brunger et al. 1990; Brunger et al. 1987; Rice and Brunger 1994). Electron density for each structure was calculated from $F_O$-$F_C$ and $2F_O$-$F_C$ using phase information averaged over the three PDB files output from simulated annealing refinement (Brunger et al. 1997; Kleywegt and Brunger 1996; Read 1986). Electron density maps were then drawn using MAPMAN (Kleywegt and Jones 1996). The $2F_O$-$F_C$ map was viewed to ensure that the peptide backbone followed the pattern of electron density. The $F_O$-$F_C$ map was then carefully analysed to identify electron density corresponding to the positions of the manganese and calcium ions and carbohydrate ligands.

Refinement and model building were then repeated cyclically to improve the model of the protein. In subsequent rounds of refinement, the rigid body refinement was omitted and no changes were made to the test-set of reflections. The temperature of simulated annealing was gradually increased to 2500K. In the latter stages of model building and refinement, water molecules were inserted using a water-pick program (Kleywegt and Brunger 1996; Read 1986).

#### 3.2.5.1 *n*ECL

Careful examination of the initial electron density maps allowed mutation of the non-conserved residues back to their native side-chains. The N- and C-terminal

residues removed from the search model were modelled into the electron density. Atomic coordinates for $Mn^{2+}$ and $Ca^{2+}$ and lactose were obtained from the Hetero-compound Information Centre (HIC-Up) (Kleywegt and Jones 1998; http://alpha2.bmc.uu.se/hicup/) and modelled into the combining site of each protomer. The geometry of the *cis*-peptide bond between $Ala^{88}$ and $Asp^{89}$ was manually built into the structure of each protomer, based upon electron density in the $F_O$-$F_C$ map. $F_O$-$F_C$ electron density revealed the positions of the oligosaccharide attached to $Asn^{113}$. Coordinates for the sugars involved were downloaded from the HIC-Up server and modelled into the electron density. As there was still some $F_O$-$F_C$ density remaining after energy minimisation and B-factor refinement, the occupancy of the sugars was reduced from 1.0 to 0.25-0.5. Water molecules were inserted into the model if there were peaks in the $F_O$-$F_C$ electron density map with heights greater than $3\sigma$ at hydrogen bond forming distances from the appropriate atoms, but those with a temperature factor higher than $60Å^2$ were excluded from the model.

#### 3.2.5.2 *rec*ECL Structures

Model building and refinement for the five *rec*ECL structures were undertaken as described for *n*ECL. Where appropriate, coordinates for combining-site ligands were obtained from the HIC-Up server and modelled into the electron density before the addition of water molecules. No oligosaccharide was modelled into the *rec*ECL structures because the protein was recombinantly over-expressed in *E.coli*, which does not have the cellular machinery to post-translationally modify proteins in that manner.

### 3.2.6 Structure Validation and Analysis

The final structural model for each ECL crystal analysed was validated using PROCHECK (Laskowski et al. 1993), a set of programs that assess the stereochemical quality of protein structures. Interactions between the $Mn^{2+}$ and $Ca^{2+}$ ions and the protein were determined using the intermolecular contacts program of the CCP4 suite (Collaborative Computational Project Number 4. 1994). In addition, the number of hydrogen bonds formed with lactose and water molecules was assessed using the hydrogen bonds program. Cross-validation of the outputs from the two programs allowed the amino acid side chains involved in carbohydrate binding to be identified. The structures of the dimers formed by *n*ECL and *rec*ECL were analysed to determine which residues are located in the dimer interface and, thus, are involved in forming protomer-protomer contacts and also to look at the influence of glycosylation on the quarternary structure. The intermolecular contacts and hydrogen bonds programs from

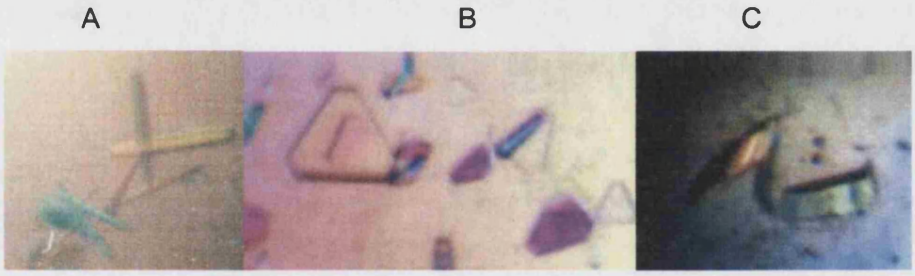
the CCP4 suite were used to identify hydrogen bonds and van der Waals contacts between two protomers in the dimer.

## 3.3 RESULTS

### 3.3.1 Crystallisation of ECL

Crystals of ECL were grown at 16°C using the hanging drop vapour diffusion method. Within 4-6 weeks, crystals were observed in three morphologies in different conditions (Figure 14): Thin, needle crystals were grown in drops equilibrated against 70% MPD and 0.1M HEPES, pH 7.5. Trigonal crystals were produced in mother liquor comprising 2.0M ammonium phosphate and 0.1M Tris-HCl, pH 8.5. A third crystal form was observed in drops equilibrated against 2.0M ammonium sulphate and 5% isopropanol.

**Figure 14 - Three Crystal Morphologies**




The initial crystallisation screens produced crystals in three different morphologies. Crystals in slide **A** were grown from mother liquor comprising 70% MPD and 0.1M HEPES. The trigonal crystals shown in slide **B** were grown from mother liquor comprising 2.0M ammonium phosphate and 0.1M Tris-HCl. Crystals in slide **C** were grown from mother liquor containing 2.0M ammonium sulphate and 5% isopropanol. These conditions were repeated and optimised to produce crystals of diffraction quality.

*n*ECL only crystallised in hexagonal form but *rec*ECL crystallised in several crystal forms under a variety of conditions. The same protein can crystallise in different space groups under different conditions because changes in pH, solvent or precipitant can alter the crystal packing (Blundell and Johnson 1976). In the case of *rec*ECL, the concentration of PEG 3350 and the presence and absence of imidazole and lactose in the crystallisation drop altered the structure of the crystal lattice, producing crystals with triclinic or monoclinic lattices. Mother liquor comprising 17% PEG 3350, 0.3M sodium chloride and 0.02M imidazole yielded crystals in space group P1. The addition of Lac to drops containing and equilibrated against the same mother liquor led to the production of crystals in space group C2. Similar crystallisation conditions also produced crystals

in space group $P2_1$. On the other hand, *rec*ECL only crystallised in trigonal form from conditions containing ammonium phosphate. *rec*ECL crystals were also grown in conditions containing ammonium sulphate and MPD, but these crystals were not of diffraction quality.

Co-crystallisation of both *n*ECL and *rec*ECL with a variety of galactose-based carbohydrates was also attempted. *n*ECL only crystallised in the presence of Lac (i.e. no crystals of *n*ECL alone were produced despite several attempts). Fortunately, *rec*ECL crystallised alone and co-crystallised with Lac, GalNAc and galactose, although crystals of the GalNAc and galactose complexes were too small to be used in X-ray diffraction experiments.

**Table 3 - Growth of ECL Crystals for Diffraction Experiments**

| Protein + Ligand | Mother Liquor | Space group | No of Molecules per Asymmetric Unit | Solvent Content (%) |
|---|---|---|---|---|
| *n*ECL + Lac | 70% MPD<br>0.1M HEPES pH 7.0 | $P6_5$ | 2 | 66.73 |
| *rec*ECL | 17% PEG 3350<br>0.3M sodium chloride<br>0.02M imidazole | P1 | 4 | 48.57 |
| *rec*ECL | 20% PEG 3350<br>0.3M sodium chloride<br>5% glycerol | $P2_1$ | 4 | 49.21 |
| *rec*ECL + Lac | 20% PEG 3350<br>0.2M imidazole | $P2_1$ | 4 | 48.17 |
| *rec*ECL + Lac | 17% PEG 3350<br>0.3M sodium chloride<br>0.02M imidazole | C2 | 3 | 64.56 |
| *rec*ECL<br>(GalNAc soak) | 2.0M ammonium phosphate<br>0.1M Tris-HCl pH 8.6<br>5% glycerol | $P3_212$ | 3 | 45.71 |

The conditions used to produce the six ECL crystals used in diffraction experiments are presented. ECL was crystallised in five different space groups under a variety of conditions. Three crystals resulted from co-crystallisation of *n*ECL and *rec*ECL with lactose (Lac) and one *rec*ECL crystal was soaked with N-acetylgalactosamine (GalNAc) prior to data collection.

A total of six ECL crystal forms in five different space groups were analysed and the conditions used to produce these crystals are detailed in Table 3. The three crystal forms of the *n*ECL-Lac and *rec*ECL-Lac complexes were produced by co-crystallisation. The three remaining crystal forms were of *rec*ECL alone, one of which

was soaked in a solution containing mother liquor and 50mM GalNAc for four hours prior to data collection. The high solvent content of protein crystals can easily be exploited to allow the diffusion of substrates, ligands or inhibitors into the crystal to interact with the protein molecules and this was the rationale behind soaking one *rec*ECL crystal with GalNAc prior to data collection.

### 3.3.2 X-Ray Diffraction Data

**Table 4 - Data Collection Strategies**

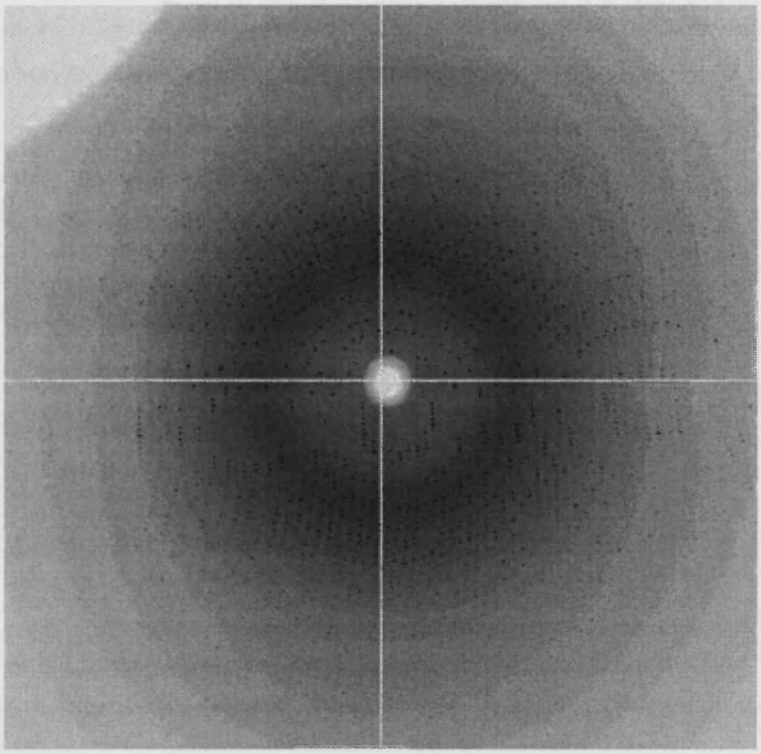
| | *n*ECL-Lac | *rec*ECL | *rec*-Lac | *rec*-Lac | *rec*ECL | *rec*ECL |
|---|---|---|---|---|---|---|
| Space group | $P6_5$ | P1 | $P2_1$ | C2 | $P3_212$ | $P2_1$ |
| Beamline | 14.1 | 14.1 | 14.1 | 14.2 | 14.2 | 14.1 |
| Distance (mm) | 100 | 110 | 75 | 200 | 200 | 70 |
| Maximum resolution (Å) | 2.00 | 2.13 | 1.70 | 2.45 | 2.30 | 1.65 |
| Oscillation range (°) | 1.0 | 1.0 | 0.4 | 1.0 | 1.0 | 0.75 |
| Exposure time (s) | 15 | 10 | 10 | 15 | 10 | 10 |
| No of images | 74 | 225 | 200 | 150 | 100 | 200 |

Data collection statistics for the six ECL crystals analysed are presented. Diffraction data were collected to high resolution using the SRS at Daresbury, UK, on beamlines 14.1 and 14.2. In each case, the data collection strategy was determined by the quality of diffraction by the crystal. A greater number of diffraction images were recorded for crystals of lower symmetry in order to collect reflections from all parts of the crystal. The crystals were stable in the X-ray beam, allowing a complete dataset to be collected from a single crystal.

Details of data collection for the crystals analysed are summarised in Table 4. A greater number of diffraction images were collected from crystals of lower order symmetry to ensure that each unique reflection was recorded at least once. The recorded diffraction patterns were well resolved (for example, Figure 15) with the intensities of the reflections easily distinguishable from the level of background radiation and lunes were clearly visible. The crystals also exhibited low mosaicity and were stable in the X-ray beam, enabling a complete dataset to be collected from a single crystal in each case. By indexing the first recorded images, the Bravais lattice of each crystal was determined.

All diffraction data for each crystal were scaled and merged until $I/\sigma I$ was greater than 2 and $\chi^2$ was close to 1 in each resolution shell. After scaling, each dataset was at least 80% complete overall (and at least 77% complete in the highest resolution shell) with low R factors. Details of diffraction data processing and refinement statistics for each of the six crystal structures are presented in Table 5.

**Figure 15 - Diffraction Pattern Recorded from a Crystal of ECL**



This diffraction pattern was recorded from the crystal of native ECL in complex with lactose (space group $P6_5$). These data were recorded at CCLRC Daresbury Laboratory on station PX14.1 ($\lambda$ = 1.488Å) to a maximum of 2.0Å resolution. The lunes and reflections are well resolved and mosaicity in the crystal is low. This diffraction pattern is typical of those recorded for the ECL crystals analysed.

The final models for each of the six structures were well refined, with values for $R_{cryst}$ and $R_{free}$ in good agreement. The average B-factors for the crystal structures were low (in the range 20.32 to 33.78Å$^2$) and the average B-factor for bulk solvent ranged from 40.38 to 59.16Å$^2$. There was low root-mean-square deviation (rmsd) from ideality for main chain bonds and bond angles. Ramachandran plots (Appendix 1) revealed that >99% of amino acid residues within each structure adopted allowed configurations and the covalent geometry (i.e. main-chain bond lengths and bond angles and planarity of aromatic rings) was within the normal range for most of the residues in each model. This is consistent with geometrically well-defined crystal structures of similar resolution. Electron density for 240 of the amino acids that comprise an ECL protomer enabled all residues except $Asn^{241}$ to be modelled. There was also well-resolved electron density for the $Ca^{2+}$ and $Mn^{2+}$ ions in all the structures and for bound carbohydrate in the ECL-Lac complexes.

**Table 5 - Data Processing and Refinement Statistics**

| | *n*ECL-Lac | *rec*ECL | *rec*ECL | *rec*-Lac | *rec*-Lac | *rec*ECL |
|---|---|---|---|---|---|---|
| **PROCESSING:** | | | | | | |
| Space group | $P6_5$ | P1 | $P2_1$ | $P2_1$ | C2 | $P3_212$ |
| Cell dimensions (Å) | a = 134.02<br>b = 134.02<br>c = 81.64<br>α = 90°<br>β = 90°<br>γ = 120° | a = 55.28<br>b = 55.37<br>c = 86.93<br>α = 86.23°<br>β = 75.37°<br>γ = 82.13° | a = 55.47<br>b = 115.17<br>c = 80.93<br>α = 90°<br>β = 93.52°<br>γ = 90° | a = 54.90<br>b = 167.23<br>c = 55.13<br>α = 90°<br>β = 97.09°<br>γ = 90° | a = 203.26<br>b = 101.13<br>c = 55.28<br>α = 90°<br>β = 102.43°<br>γ = 90° | a = 81.05<br>b = 81.05<br>c = 190.95<br>α = 90°<br>β = 90°<br>γ = 120° |
| Vm ($Å^3$/Da) | 3.73 | 2.45 | 2.48 | 2.41 | 3.53 | 2.32 |
| Resolution (Å) | 2.00 | 2.13 | 1.65 | 1.70 | 2.45 | 2.13 |
| Reflections measured | 414754 | 399699 | 306951 | 1021770 | 440979 | 362378 |
| Unique reflections | 56451 | 55200 | 109371 | 108695 | 40576 | 32363 |
| Completeness (%) (outermost shell) | 99.7<br>(99.9) | 93.1<br>(86.1) | 89.9<br>(88.9) | 94.1<br>(77.3) | 81.0<br>(82.1) | 99.9<br>(100.0) |
| I/σI (outermost shell) | 18.44<br>(6.22) | 10.93<br>(5.29) | 11.57<br>(2.03) | 28.05<br>(5.26) | 5.83<br>(1.66) | 15.5<br>(3.55) |
| $R_{symm}$ * | 0.070 | 0.077 | 0.085 | 0.058 | 0.122 | 0.135 |
| **REFINEMENT:** | | | | | | |
| $R_{cryst}$ (%) ** | 21.14 | 19.29 | 25.52 | 17.79 | 25.58 | 21.85 |
| $R_{free}$ (%) ** | 24.70 | 25.36 | 29.03 | 17.84 | 33.29 | 29.35 |
| Average B factor ($Å^2$) | 27.96 | 24.45 | 23.50 | 20.32 | 33.78 | 28.60 |
| Wilson B factor ($Å^2$) | 23.1 | 25.5 | 20.8 | 19.2 | 40.5 | 32.4 |
| RMSD from ideality: | | | | | | |
| Bonds (Å) | 0.006 | 0.006 | 0.006 | 0.008 | 0.008 | 0.008 |
| Bond angles (deg) | 1.42 | 1.40 | 1.41 | 1.49 | 1.49 | 1.65 |
| Dihedrals (deg) | 25.32 | 25.66 | 25.67 | 25.67 | 25.64 | 25.89 |
| Impropers (deg) | 0.77 | 0.75 | 0.84 | 0.80 | 0.89 | 0.83 |

* $R_{symm} = \Sigma_{hkl}\Sigma_i|I_i(hkl)-\langle I\rangle(hkl)| / \Sigma_{hkl}\Sigma_i I_i(hkl)$, where $\langle I\rangle$ is the averaged intensity of the *i* observations of reflection hkl

** $R_{cryst} = \Sigma||F_O|-|F_C|| / \Sigma|F_O|$, where $F_O$ and $F_C$ are the observed and calculated structure factor amplitudes, respectively

** $R_{free}$ is equal to $R_{cryst}$ for a randomly selected 5% subset of reflections not used in the refinement

Data processing and refinement statistics for the six crystal structures are presented. Data were collected to high resolution and at least 75% completeness. In each case, I/σI was greater than 2.0 and $R_{symm}$ was low. The final models were well refined with low B factors and values for $R_{cryst}$ and $R_{free}$ in good agreement.

### 3.3.2.1 Native ECL in Complex with Lactose

*n*ECL co-crystallised with Lac in space group $P6_5$, producing small, needle-like crystals that diffracted to a maximum of 2.0Å resolution. There were two *n*ECL molecules per asymmetric unit and 67% of the crystal volume was occupied by solvent. Indexing revealed the Bravais lattice of the crystal to be hexagonal and scaling reduced the possible space groups to $P6_1$ and $P6_5$. Molecular replacement gave the correct space group as $P6_5$ (Table 6) identified by the highest correlation coefficient and lowest R factor.

**Table 6 - Determination of *n*ECL Space Group**

| | $P6_1$ | | $P6_5$ | |
|---|---|---|---|---|
| | CCF | RF | CCF | RF |
| Roting | 17.4 | 55.7 | 17.4 | 55.7 |
| Traing | 22.6 | 54.1 | 36.4 | 49.7 |
| Traing2 | 31.2 | 51.0 | 59.7 | 40.6 |
| Fiting | 42.7 | 48.2 | 73.3 | 34.4 |

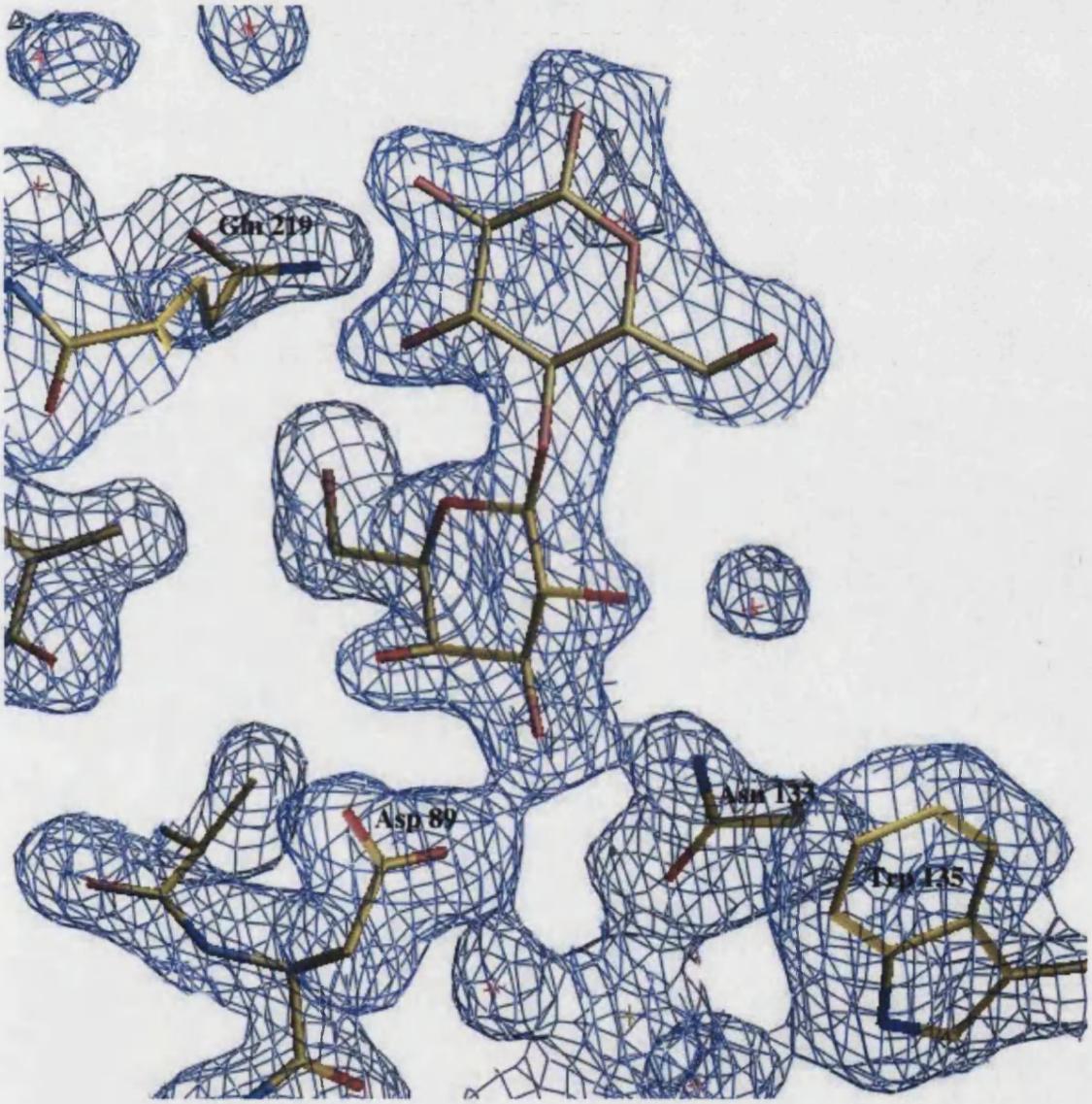
The results of molecular replacement for *n*ECL in the four hexagonal space groups $P6_1$, $P6_5$, $P6_122$ and $P6_522$ are presented. CCF and RF are the correlation coefficient and R-factor, respectively, of the rotation function (Roting), translation functions (Traing and Traing2) and the refinement function (Fiting) of the molecular replacement program AMoRe (Navaza 1994). The correct space group was easily identified as $P6_5$ because it was the space group that gave the highest correlation coefficient and lowest R factor in each of the rotation and translation searches.

The initial electron density maps for the *n*ECL-Lac complex were well defined with few breaks in the density. The peptide backbone of both protein molecules in the asymmetric unit fitted well into the observed $2F_O$-$F_C$ electron density. One lactose molecule was bound in the combining site of each *n*ECL molecule, with $F_O$-$F_C$ electron density for the disaccharide clearly resolved (Figure 16). Lactose was modelled into the combining site and made hydrogen bond contacts with combining site residues through structural water molecules (described in section 3.3.3.3). There were also clear peaks in the $F_O$-$F_C$ electron density map close to the carbohydrate-combining site corresponding to the calcium and manganese ions and their associated water molecules.

There was also some $F_O$-$F_C$ electron density surrounding the peptide bond between $Ala^{88}$ and $Asp^{89}$ and analysis of the Ramachandran plot revealed that the side chain of $Asp^{89}$ adopted a disallowed configuration. This was because the peptide bond was originally modelled in the *trans* conformation when, in fact, this peptide bond

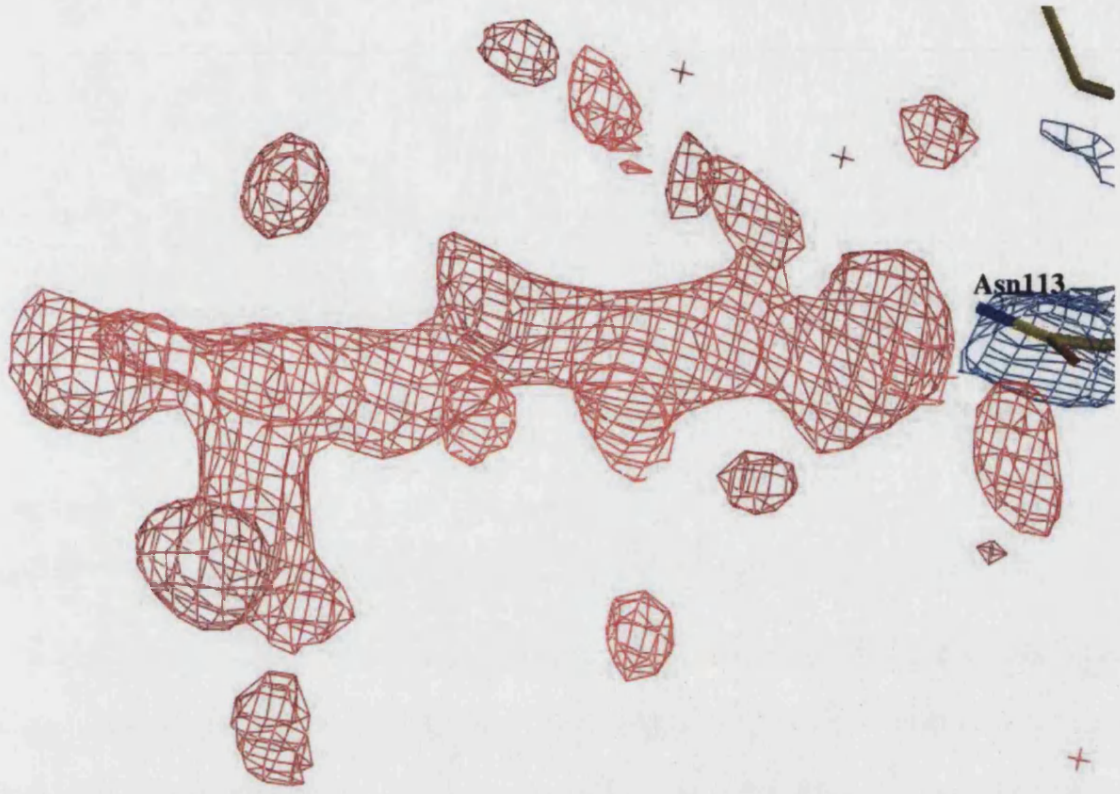
should be in the *cis* conformation. There is a *cis*-peptide bond at this position in all of the legume lectin structures solved to date (Lis and Sharon 1998) that is proposed to hold the side chain of $Asp^{89}$ in the correct orientation for carbohydrate binding. The *cis*-configuration of the peptide bond is stabilised by one of the water molecules that coordinates the calcium ion.

**Figure 16 - Electron Density for Lactose in the Combining Site of *n*ECL**




The initial electron density maps clearly defined the location of the lactose molecule within the combining site of each *n*ECL molecule in the structural model. The disaccharide was modelled into the electron density, as shown above, such that the galactose moiety was in close proximity to combining site residues $Asp^{89}$ and $Asn^{133}$. Lactose fitted well into the observed electron density and made hydrogen bond contacts with a number of combining site residues through structural water molecules (see Figure 25).

**Figure 17 - $F_O$-$F_C$ Electron Density For the Glycosylation Attached to $Asn^{113}$**




The glycosylation profile of ECL has been characterised and the major component is a heptasaccharide (Ashford et al. 1991). The same glycosylation profile has been observed in other lectins isolated from *Erythrina* species and also in some unrelated legume lectins. This figure shows the $F_O$-$F_C$ electron density observed near the side chain of $Asn^{113}$ in one of the *n*ECL molecules in the structural model, indicating the positions of the first three sugar residues of the attached oligosaccharide. The $F_O$-$F_C$ density observed for the oligosaccharide in the second protein molecule was much stronger, enabling six of the seven sugar residues to be modelled (see Figure 21).

Further analysis of the initial electron density maps revealed a large region of $F_O$-$F_C$ electron density near the side chain of $Asn^{113}$ in both of the protein molecules (Figure 17). Since $Asn^{113}$ is the only potential glycosylation site in the *n*ECL sequence (Figure 13), this electron density was assumed to represent the bound oligosaccharide. The glycosylation profile of ECL has been characterised (Ashford et al. 1991) and the heptasaccharide α-D-Man*p*-(1→3)-[α-D-Man*p*-(1→6)]-[β-D-Xyl*p*-(1→2)]-β-D-Man*p*-β-D-Glc*p*NAc-(1→4)-[α-L-Fuc*p*(1→3)]-D-GlcNAc – where Xyl is xylose – makes up the major component. Residues from this profile were modelled into the observed electron density.

The final structural model for *n*ECL had an $R_{cryst}$ of 21.14% ($R_{free}$ 24.70%) and contained 476 amino acids, two lactose molecules, two calcium and manganese ions,

413 water molecules, four GlcNAc residues and two Fuc, one Xyl and two Man residues. The average B factor for protein atoms was 27.96Å$^2$ compared to 59.16Å$^2$ for bulk solvent and the estimated Luzzati coordinate error was 0.24Å. Analysis of the Ramachandran plot revealed that only two residues, Tyr$^{106}$ from each molecule, adopted generously allowed configurations (all other residues had allowed and additional allowed configurations). The side chain of only one of the aromatic residues was non-planar and a potential alternative side-chain configuration was observed for Arg$^{50}$, Ile$^{59}$ and Gln$^{202}$ in one of the protein molecules in the asymmetric unit.

### 3.3.2.2 Recombinant ECL

*rec*ECL was crystallised in the absence of any added carbohydrate ligand in two crystal forms. The triclinic form had four molecules per unit cell with 49% of the crystal occupied by solvent and the monoclinic form also had four molecules per asymmetric unit with 49% of the crystal volume occupied by solvent.

X-ray diffraction data for *rec*ECL in triclinic form were recorded to a maximum of 2.13Å resolution. $F_O$-$F_C$ density observed in the combining sites of two of the protein molecules was modelled as glycerol because glycerol was added to the cryoprotectant solution used prior to data collection. The final model had an $R_{cryst}$ of 19.29% ($R_{free}$ 25.36%) and contained 958 amino acids, four calcium and manganese ions, three glycerol molecules and 856 water molecules. The average B factor was 24.45Å$^2$ for protein atoms and 58.98Å$^2$ for bulk solvent. The estimated Luzzati coordinate error was 0.23Å. The Ramachandran plot revealed no amino acid side chains were in disallowed configurations and that only one residue (Tyr$^{106}$ in one protein molecule) adopted a generously allowed configuration. Possible alternative configurations were observed in the $F_O$-$F_C$ electron density map for the side chains of Arg$^{50}$, Arg$^{84}$, Met$^{95}$ and Ser$^{120}$, but not in all four of the protein molecules in the unit cell.

Data for the monoclinic form of *rec*ECL (space group $P2_1$) were recorded to 1.65Å resolution. Although no carbohydrate had been added to the crystallisation drop, $F_O$-$F_C$ electron density corresponding to the structure of two pyranose sugar rings was observed in the combining site of three of the four molecules. Lac was modelled into this $F_O$-$F_C$ density and the R factors dropped in subsequent refinement steps. The final model of *rec*ECL in space group $P2_1$ had an $R_{cryst}$ of 25.52% ($R_{free}$ 29.03%) and contained 953 amino acids, three Lac molecules, four calcium and manganese ions and 745 water molecules. The average B factor was 23.50Å$^2$ for protein atoms and 53.93Å$^2$ for bulk solvent. The estimated Luzzati coordinate error was 0.28Å. No amino acid side chains were located in disallowed regions of the Ramachandran plot,

although six side chains ($Tyr^{106}$ in all four molecules and $Leu^{109}$ and $Asp^{221}$ each in only one molecule) adopted generously allowed configurations. Possible alternative configurations were noted for the side chains of residues $Leu^{20}$, $Ser^{29}$, $Leu^{34}$, $Ile^{59}$, $Val^{67}$, $Arg^{84}$, $Met^{95}$, $Lys^{116}$, $Ile^{197}$ and $Leu^{237}$ in at least one of the four protein molecules present in the crystal structure.

Investigation into why Lac was present in the combining site of *rec*ECL when no carbohydrate had been added to the crystallisation drop revealed that Lac was likely present as a result of the purification procedure. The final purification step involves an immobilised-lactose affinity column from which *rec*ECL is eluted with 0.3M lactose. Although *rec*ECL was dialysed after purification, analysis by mass spectroscopy revealed a shoulder on the ECL peak (26.7kDa) of higher molecular weight (27.5kDa), that could represent one ECL protomer bound to two lactose molecules (J.A.Chaddock, personal communication). As ECL exhibits a high binding affinity for Lac – 4 units of the lectin are completely inhibited by 2.0mM Lac (Iglesias et al. 1982) – it is not surprising that dialysis of purified *rec*ECL was not sufficient to remove all the bound lactose. However, the peak in the mass spectrum postulated to represent an ECL-Lac complex was small and not significantly higher than the background signal. This low level of contamination might explain why not all of the protein molecules in the affected structures were seen to bind lactose. Since an electron density map represents a time-averaged structure, if not all the lectin protomers in the crystal bound Lac it would not be observed in all the protein molecules in the structural model. This could also explain why Lac was not observed in the triclinic structure of *rec*ECL, even though it was crystallised from the same batch of protein. Unfortunately, the presence of Lac in the protein sample was not detected until the electron density maps were analysed.

### 3.3.2.3 Recombinant ECL Complexes

In addition, three *rec*ECL-carbohydrate complexes were studied. Two were monoclinic crystals, space groups $P2_1$ and C2, produced by co-crystallising *rec*ECL with Lac. The third complex in trigonal form (space group $P3_212$) was obtained by soaking one *rec*ECL crystal in a solution containing 50mM GalNAc for four hours prior to data collection.

*rec*ECL co-crystallised with Lac in space group $P2_1$ with four molecules per asymmetric unit and 49% of the crystal volume occupied by solvent. Each of the four protein molecules had Lac bound at the combining site. The final model at 1.70Å resolution had an $R_{cryst}$ of 17.79% ($R_{free}$ 17.84%) and contained 958 amino acids, four Lac molecules, four calcium and manganese ions and 1089 water molecules. The

average B factor was 20.32Å$^2$ for protein atoms and 52.73Å$^2$ for bulk solvent and the estimated Luzzati coordinate error was 0.18Å. One main chain bond was considered too short – Cα-Cβ in Asp$^{89}$, which takes part in the conserved *cis*-peptide bond. The Ramachandran plot revealed that five amino acid side chains adopted generously allowed configurations (Tyr$^{106}$ in all four protomers and Asp$^{221}$ in only one protomer) but no side chains had disallowed configurations. The following amino acid side chains were found to have possible alternative configurations, but not in all of the protein molecules in the crystal structure: Thr$^3$, Asn$^{16}$, Ile$^{25}$, Met$^{42}$, Thr$^{54}$, Lys$^{55}$, Ile$^{59}$, Ser$^{77}$, Arg$^{84}$, Met$^{95}$, Lys$^{99}$, Gln$^{104}$, Asn$^{114}$, Asn$^{119}$, Ser$^{120}$, Ser$^{152}$, Ser$^{175}$, Leu$^{180}$, Val$^{182}$, Arg$^{220}$, His$^{234}$ and Ser$^{236}$.

*rec*ECL also co-crystallised with Lac in space group C2 with three molecules per asymmetric unit and 65% of the crystal volume occupied by solvent. However, Lac could only be modelled into $F_O$-$F_C$ density in the combining sites of two of the three molecules. The final model had an $R_{cryst}$ of 25.58% ($R_{free}$ 33.29%) and contained 712 amino acids, two Lac molecules, three calcium and manganese ions and 218 water molecules. The estimated Luzzati coordinate error was 0.38Å. The average B factor was 33.78Å$^2$ for protein atoms and 47.53Å$^2$ for bulk solvent. Analysis of the Ramachandran plot revealed that no residues adopted disallowed configurations, but four side chains had generously allowed configurations of the plot – Tyr$^{106}$ in each molecule and Ile$^4$ from one molecule. Only residue Arg$^{84}$ in one of the three protein molecules was observed to have a potential alternative configuration.

One crystal of *rec*ECL in trigonal space group $P3_212$ was soaked with GalNAc because co-crystallisation had been unsuccessful. There were three molecules per asymmetric unit with 46% of the crystal volume occupied by solvent. Although the Bravais lattice was identified as trigonal by indexing, the space group had to be uniquely identified by molecular replacement, as described for *n*ECL. Scaling reduced the number of possible space groups to $P3_121$, $P3_112$ and $P3_212$ and was confirmed as $P3_212$ by molecular replacement (Table 7). Initially, GalNAc was modelled into the $F_O$-$F_C$ electron density observed in the combining site of each molecule but during later stages of refinement it became clear that there was extra density consistent with the presence of another pyranose sugar ring. Lac was then modelled into the density and found to fit neatly. This trigonal crystal was grown from the same batch of protein as that used to grow the monoclinic *rec*ECL crystal that unexpectedly had lactose present in the combining site (section 3.3.2.2). As this batch of protein was contaminated with lactose, soaking with GalNAc would not provide any data on a new complex structure because lactose is a stronger inhibitor of the activity of this lectin (Iglesias et al. 1982).

Only N-acetyllactosamine, fucosyllactose and fucosyllactosamine bind ECL with higher affinity than lactose (Iglesias et al. 1982; Teneberg et al. 1994).

**Table 7 - Determining the Space Group of *rec*ECL in Trigonal Crystal Form**

| | $P3_121$ | | $P3_112$ | | $P3_212$ | |
|---|---|---|---|---|---|---|
| | CCF | RF | CCF | RF | CCF | RF |
| Roting | 13.0 | 57.4 | 13.1 | 57.3 | 13.1 | 57.3 |
| Traing | 16.7 | 57.2 | 16.9 | 57.2 | 22.3 | 55.2 |
| Traing2 | 19.0 | 57.3 | 18.5 | 56.6 | 33.8 | 51.5 |
| Fiting | 33.7 | 54.4 | 30.3 | 54.8 | 48.6 | 47.6 |

The space group could not be uniquely identified without performing the rotation and translation searches during molecular replacement, as previously described for *n*ECL in complex with lactose. The calculated values for the correlation coefficient (CCF) and R-factor (RF) for the top peak in each search revealed the space group to be $P3_212$.

The final model of *rec*ECL in trigonal form at 2.13Å resolution had an $R_{cryst}$ of 21.85% ($R_{free}$ 29.35%) and contained 713 amino acids, three Lac molecules, three calcium and manganese ions and 412 water molecules. The average B factor for protein atoms was 28.60Å$^2$ and 40.38Å$^2$ for bulk solvent. The estimated Luzzati coordinate error was 0.29Å. Analysis of the Ramachandran plot revealed that four residues were located in generously allowed regions ($Tyr^{106}$ in two molecules and $Asp^{221}$ and $Asp^{118}$ each in one molecule). No residues were found in disallowed regions of the plot. Seven aromatic side chains were found to deviate from planarity. The side chains of residues $Arg^{50}$, $Val^{67}$, $Ala^{68}$, $Ser^{77}$, $Ala^{88}$, $Gln^{117}$, $Gln^{159}$, $Leu^{180}$ and $Lys^{201}$ were observed to have possible alternative configurations, but not in all three of the protein molecules in the asymmetric unit.

### 3.3.3 Crystal Structure of ECL

#### 3.3.3.1 Tertiary Structure

Protomers of *n*ECL and *rec*ECL adopt the jelly-roll fold that is characteristic of the legume lectins (Figure 18). As described previously, the legume lectin fold is made up of a flat, six-stranded β-sheet; a curved, seven-stranded β-sheet; a short five-membered β-sheet and several loops that connect the three sheets. The seven- and six-stranded sheets are both entirely anti-parallel.

Between the six ECL structures solved, 240 of the 241 amino acids that make up one protomer could be modelled into the observed electron density. Only residue $Asn^{241}$ could not be located in any of the structural models. When determining protein

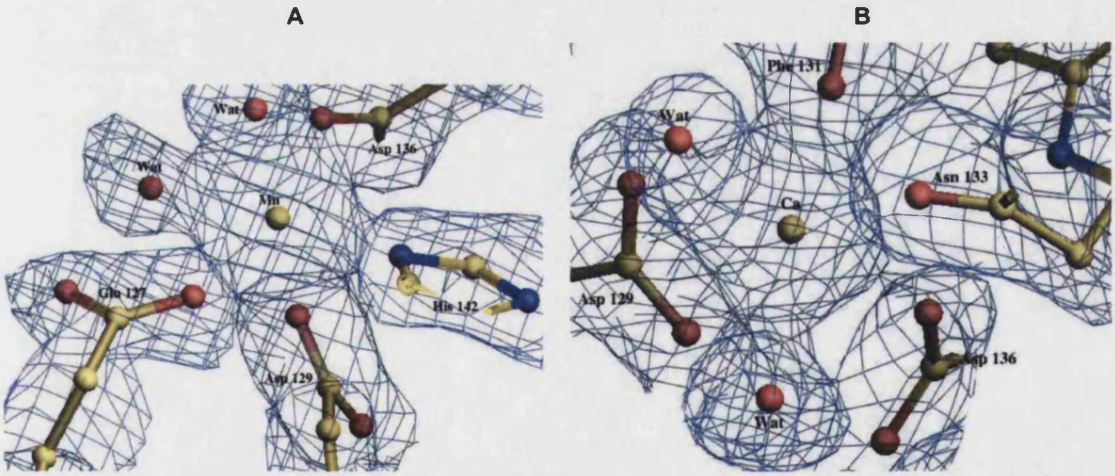
structures by X-ray crystallography it is not always possible to model residues at the N- and C-termini because these regions can be relatively mobile. In the case of ECL, the N- and C-termini associate together, forming the first two strands of the six-stranded β-sheet and, consequently, the electron density for these residues was well resolved in the six structural models. However, not all 240 residues were modelled in all of the protein molecules in all six structural models.

**Figure 18 - Jelly-Roll Structure of the ECL Protomer**


The tertiary structure of *n*ECL and *rec*ECL adopts the same jelly-roll fold in common with all other legume lectins that have been structurally characterised. This motif comprises a flat, six-stranded "back" β-sheet, a curved, seven-stranded "front" β-sheet and a short five-stranded β-sheet with loops connecting the three sheets. The calcium and manganese ions (shown as green and navy spheres, respectively) are bound close to the combining site. This figure was prepared using Molscript, Povray and GIMP.

The combining site is located in a shallow cleft on the surface of the lectin and accommodates only the galactose moiety of bound carbohydrates. Analysis of the crystal structures of other galactose-specific legume lectins - ECorL (Shaanan et al. 1991); SBA (Dessen et al. 1995); GSIV (Delbaere et al. 1993); and PNA (Banerjee et al. 1994), revealed that the conserved aspartic acid donated by combining site loop A recognises the equatorial hydroxyl of C3 and axial hydroxyl of C4 in the galactose ring (Sharma and Surolia 1997). The C3 hydroxyl group also forms hydrogen bonds with the backbone amine of the conserved glycine residue in loop B and the asparagine in loop C. The C4 hydroxyl group hydrogen bonds with residues in loop D. The corresponding residues in ECL that can be expected to be involved in carbohydrate binding are: $Asp^{89}$, $Gly^{107}$, $Phe^{131}$ and $Asn^{133}$. The differing affinities of these lectins for galactose and GalNAc have been attributed to the length of combining site loop C (i.e. those lectins with a shorter loop exhibit a preference for GalNAc and those with a longer loop bind galactose with higher affinity) (Sharma and Surolia 1997). The mode of carbohydrate binding by ECL is discussed later in section 3.3.3.3.

**Figure 19 - Electron Density for $Ca^{2+}$ and $Mn^{2+}$ Ions In the Combining Site of ECL**




In the figures above, the metal ions are depicted as yellow spheres and water molecules as red spheres. Panel A (left) shows electron density for the four amino acid side chains and two water molecules that coordinate the $Mn^{2+}$ ion bound near the combining site of ECL. Panel B (right) shows electron density for the waters and side chains that coordinate the $Ca^{2+}$ ion. Both metal ions are required for the carbohydrate binding activity of ECL and are thought to stabilise the spatial arrangement of key combining site residues.

In common with all legume lectins, each ECL protomer contains one calcium and one manganese ion, both of which are located close to the combining site (Figure 18). Their presence in the structure is associated with maintaining the correct spatial

orientation of the combining site residues for carbohydrate binding. Treatment of ECL with chelating agents did not remove the metal ions from the lectin or ablate its carbohydrate-binding activity (Iglesias et al. 1982). $Ca^{2+}$ and $Mn^{2+}$ are situated approximately 4.3Å apart in the ECL protomer. As shown in Figure 19 and detailed in Table 8, each metal coordinates two water molecules and makes contacts with four amino acid side chains. $Mn^{2+}$ makes contacts with $Glu^{127}$, $Asp^{129}$, $Asp^{136}$ and $His^{142}$ and $Ca^{2+}$ makes contacts with $Asp^{129}$, $Phe^{131}$, $Asn^{133}$ and $Asp^{136}$. One of the water molecules that co-ordinates $Ca^{2+}$ also stabilises the conserved *cis*-peptide bond ($Ala^{88}$-$Asp^{89}$) that correctly orients the side chain of $Asp^{89}$ for carbohydrate binding. In addition to binding $Ca^{2+}$, the side chains of residues $Phe^{131}$ and $Asn^{133}$ are also implicated in lactose binding.

**Table 8 – Metal Ion Binding in ECL**

| Metal Ion | Amino Acid | Distance (Å) |
|---|---|---|
| $Ca^{2+}$ | $Asp^{129}$ CG | 2.9 |
| | $Asp^{129}$ OD1 | 2.6 |
| | $Asp^{129}$ OD2 | 2.5 |
| | $Phe^{131}$ O | 2.5 |
| | $Asn^{133}$ OD1 | 2.5 |
| | $Asp^{136}$ OD2 | 2.5 |
| $Mn^{2+}$ | $Glu^{127}$ OE2 | 2.2 |
| | $Asp^{129}$ CG | 3.2 |
| | $Asp^{129}$ OD2 | 2.2 |
| | $Asp^{136}$ CG | 3.1 |
| | $Asp^{136}$ OD1 | 2.2 |
| | $His^{142}$ NE2 | 2.4 |
| | $His^{142}$ CE1 | 3.2 |

This table details the contact distances between amino acid residues within and near the combining site of ECL and the bound calcium and manganese ions.

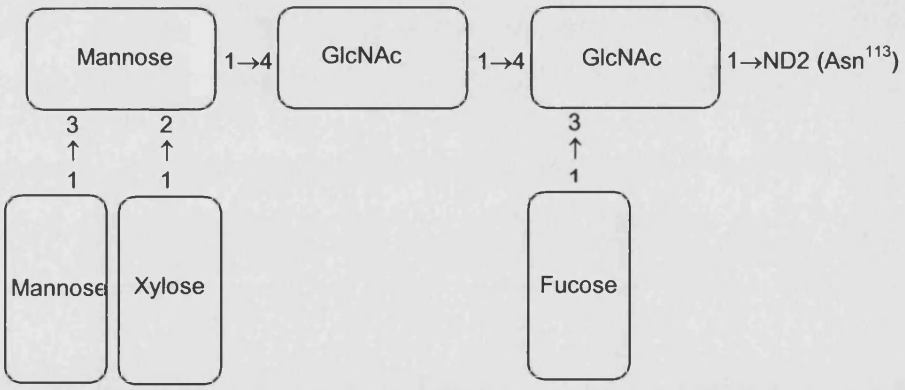
Analysis of the six ECL structures revealed a number of residues with possible alternative side chain configurations. Although these amino acids were not observed to have alternative configurations in all of the protomers in all of the structures, several of the residues ($Arg^{50}$, $Val^{67}$, $Ser^{77}$, $Arg^{84}$, $Met^{95}$, $Ser^{120}$ and $Leu^{180}$) were noted to have possible alternative configurations in more than one form of ECL. Of these seven amino acids, only $Arg^{84}$ is involved in its biological activity as one of the residues that stabilise the dimer interface (see section 3.3.3.4). It is possible that the long side chains of arginine and methionine may be flexible in the crystal structure and therefore adopt

more than one configuration. $F_O$-$F_C$ density associated with the side chains of serine residues at positions 77 and 120 may represent alternative configurations or indicate the presence of a water molecule interacting with the hydroxyl group. Since valine and leucine are aliphatic in nature, water molecules will not be associated with their side chains. Instead, it is more likely that the observed $F_O$-$F_C$ electron density represents a small amount of movement in the positions of side chain atoms.

The crystal structures of *n*ECL and *rec*ECL at 2.0Å and 2.13Å, respectively, superimpose with an overall rmsd of 0.31Å. Comparison of the structures indicated that there are no significant differences between the crystal structures of native and recombinant forms of ECL and, thus, confirmed that *rec*ECL adopts a native structure. The only difference between native and recombinant forms of ECL is the absence of N-linked carbohydrate from *rec*ECL. This is a consequence of using *E.coli* as an expression host. Therefore, it can be concluded that the presence of bound oligosaccharide does not influence the tertiary structure of ECL.

### 3.3.3.2 N-Linked Glycosylation

**Figure 20 - Glycosylation Profile of *n*ECL**




This cartoon represents the sugar residues and glycosidic linkages in the hexasaccharide modelled into the structure of *n*ECL. The major component of the glycosylation on ECL is a heptasaccharide. Six of the seven sugar residues were observed in one of the protein molecules, but weaker electron density in the other protein molecule allowed only the first three residues to be modelled.

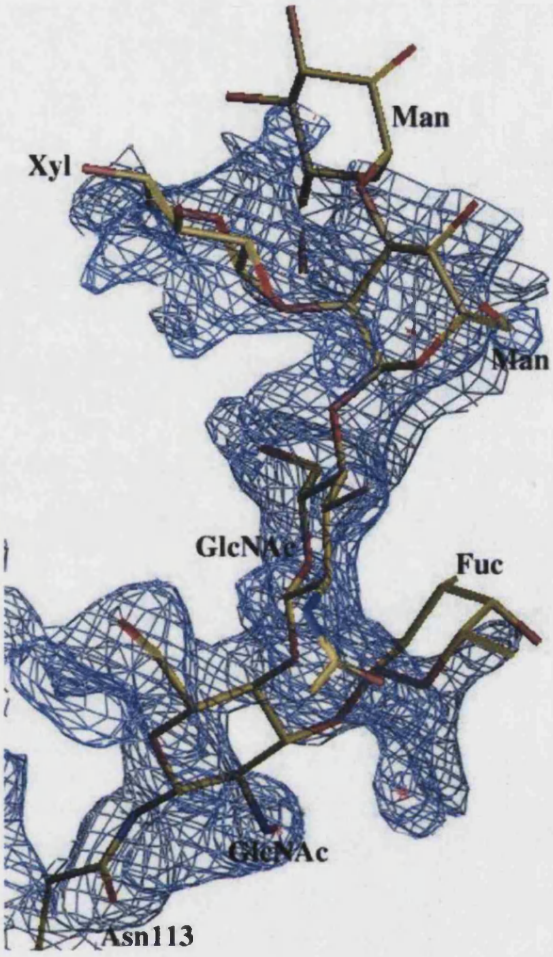
The oligosaccharide bound to *n*ECL is covalently linked to ND2 of amino acid residue $Asn^{113}$. This residue is the only possible location for the attachment of N-linked glycosylation in the 241-amino acid sequence of ECL (Figure 13). The initial $F_O$-$F_C$ electron density map for *n*ECL revealed the positions of several sugar residues (Figure 17). In one of the two protein molecules in the structural model, there was enough

electron density to model six of the seven sugar residues in the heptasaccharide that makes up the major component of carbohydrates bound to ECL. This hexasaccharide has the profile displayed in Figure 20. The trisaccharide modelled into the other protein molecule comprised only the first three residues of this hexasaccharide.

The modelled sugars fitted well into the observed electron density (Figure 21). However, after energy minimisation and B-factor refinement there was still some $F_O$-$F_C$ density surrounding the oligosaccharide in each protein molecule of the structural model. This residual $F_O$-$F_C$ density likely reflects the flexible structure of the bound carbohydrate, which can be expected to be mobile in the crystal because it is present on the surface of the lectin and is oriented away from the body of the protein. The structural flexibility and inherent mobility of the oligosaccharide produces a diffuse electron density profile because the electron density map represents a time-averaged picture of all structures in the crystal. Therefore, if the carbohydrate occupies a slightly different position in each protein molecule, its exact orientation and location cannot be pinpointed because the electron density map will reveal the positions of all the sugars in the crystal. To reduce the residual $F_O$-$F_C$ density, the occupancy of the modelled sugars was reduced from 1.0 to 0.25 for the trisaccharide in one molecule and 0.5 for the hexasaccharide in the other. In a good structural model, all atoms will have occupancy of 1.0, indicating that their positions are accurately modelled and represent their true positions. Reducing the occupancy of atoms indicates that they are not accurately positioned for all of the molecules in the crystal structure.

The structural flexibility and mobility of oligosaccharides also explains why weaker $F_O$-$F_C$ density was observed for the carbohydrate in one of the molecules in the structural model compared to the other. However, several oligosaccharides have been characterised from ECL (Ashford et al. 1991). Whilst the major component is known to be the heptasaccharide represented in Figure 12, three smaller carbohydrates, including a hexasaccharide, have also been identified. The structure of the hexasaccharide differs from the major component by the absence of the fucose residue, therefore it was obvious from the observed electron density that the bound oligosaccharide was in fact the major component but that there was no density recorded for the seventh residue. The fact that the major heptasaccharide component makes up 79% of the carbohydrate bound to ECL (Ashford et al. 1991), coupled with the observation that ECL runs as two bands on SDS-PAGE (Iglesias et al. 1982) indicates that ECL exists as a heterodimer in which the two protomers are differently glycosylated.

**Figure 21 - Electron Density for the Hexasaccharide Modelled into the Structure of *n*ECL**

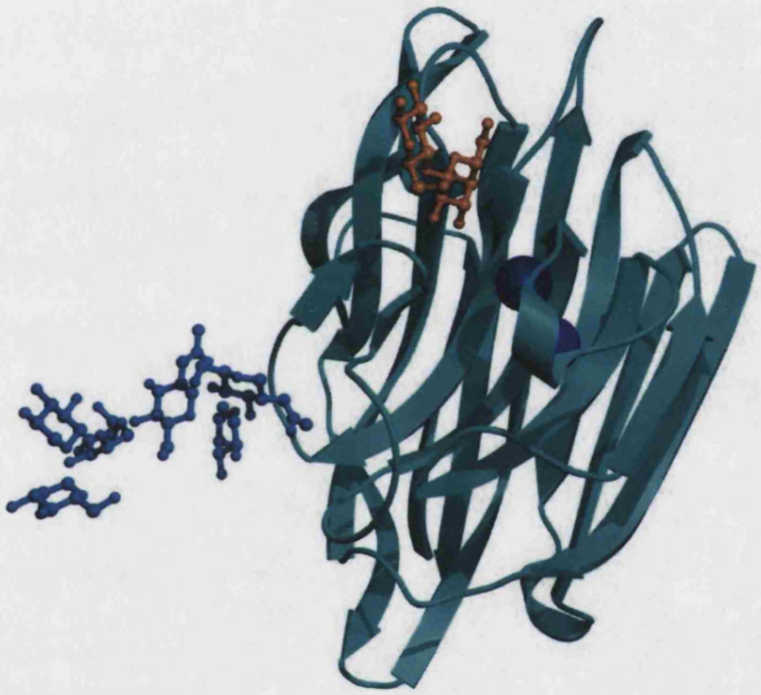



This figure shows how the six sugar residues were modelled into the electron density for one of the *n*ECL protomers in the crystal structure. Unfortunately, no electron density was observed for the seventh sugar residue (another mannose) so it could not be modelled into the structure. There was some $F_O$-$F_C$ density associated with the mannose and xylose residues and this was probably because of the structural flexibility of the bound oligosaccharide.

As shown in Figure 22, $Asn^{113}$ is located in a loop associated with the curved, seven-stranded β-sheet of the legume lectin fold. The carbohydrate-combining site is also associated with loops on this face of the lectin, but the combining and glycosylation sites are not adjacent in the ECL protomer. The bound oligosaccharide does not influence the carbohydrate-binding properties of ECL. This is demonstrated by the equivalent specificities and affinities of *n*ECL and *rec*ECL for carbohydrates and DRG (Stancombe et al. 2003). The oligosaccharide is also not located close to the dimer interface and plays no part in the quarternary association of ECL protomers (discussed later in section 3.3.3.4). Analysis of the modelled hexasaccharide revealed

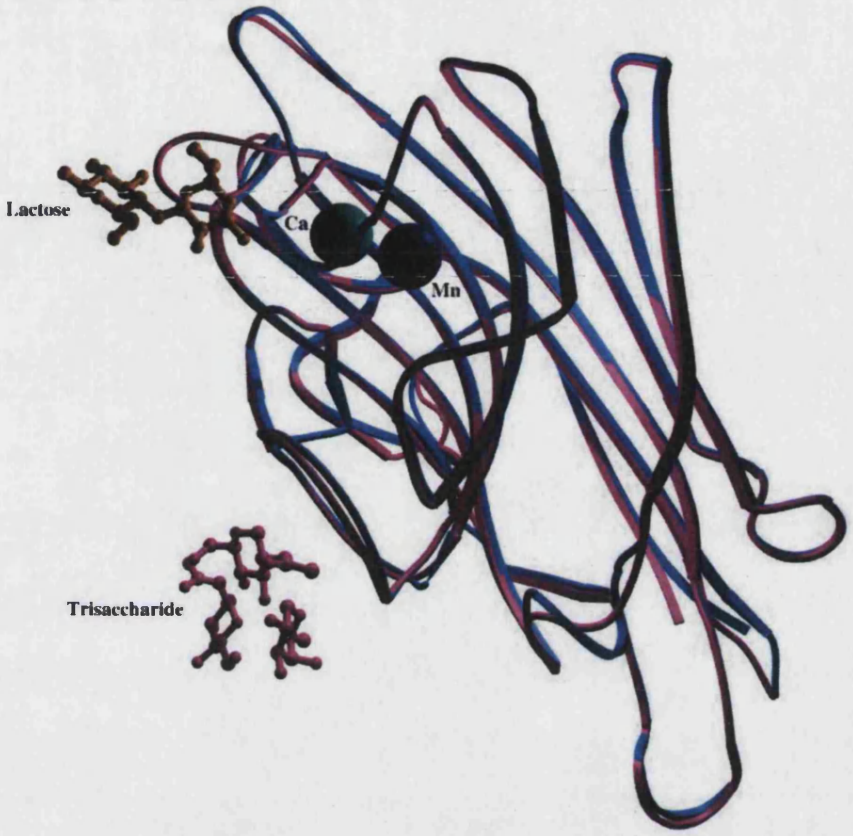
that whilst the sugars in the bound oligosaccharide make contacts with each other, with $Asn^{113}$ and with water molecules, they do not make contacts with any other parts of the lectin protomer. The role of glycosylation on legume lectins is currently unknown, but the sugar might act as a recognition motif for other proteins. ECL is synthesised with a signal sequence and is post-translationally modified. It is possible that glycosylation of ECL targets the lectin to a specific cellular location. However, until the biological function of ECL in the legume is determined, the role of bound carbohydrate remains the subject of speculation.

**Figure 22 - Location of the Glycosylation Site Relative to the Combining Site**



This figure, prepared using Molscript, Povray and GIMP, displays the location of the bound oligosaccharide in ECL with respect to the combining site. The hexasaccharide (depicted in blue) is attached to $Asn^{113}$, which is located in a loop associated with the concave face of the seven-stranded β-sheet. The combining site is a shallow cleft formed by four loops also associated with the seven-stranded β-sheet. It can be seen that the glycosylation site and combining site are not adjacent in the tertiary structure of ECL and that there are no contacts between the oligosaccharide and the bound ligand. The oligosaccharide is also not associated with the dimer interface and, with the exception of $Asn^{113}$, does not make contacts with any part of the lectin.

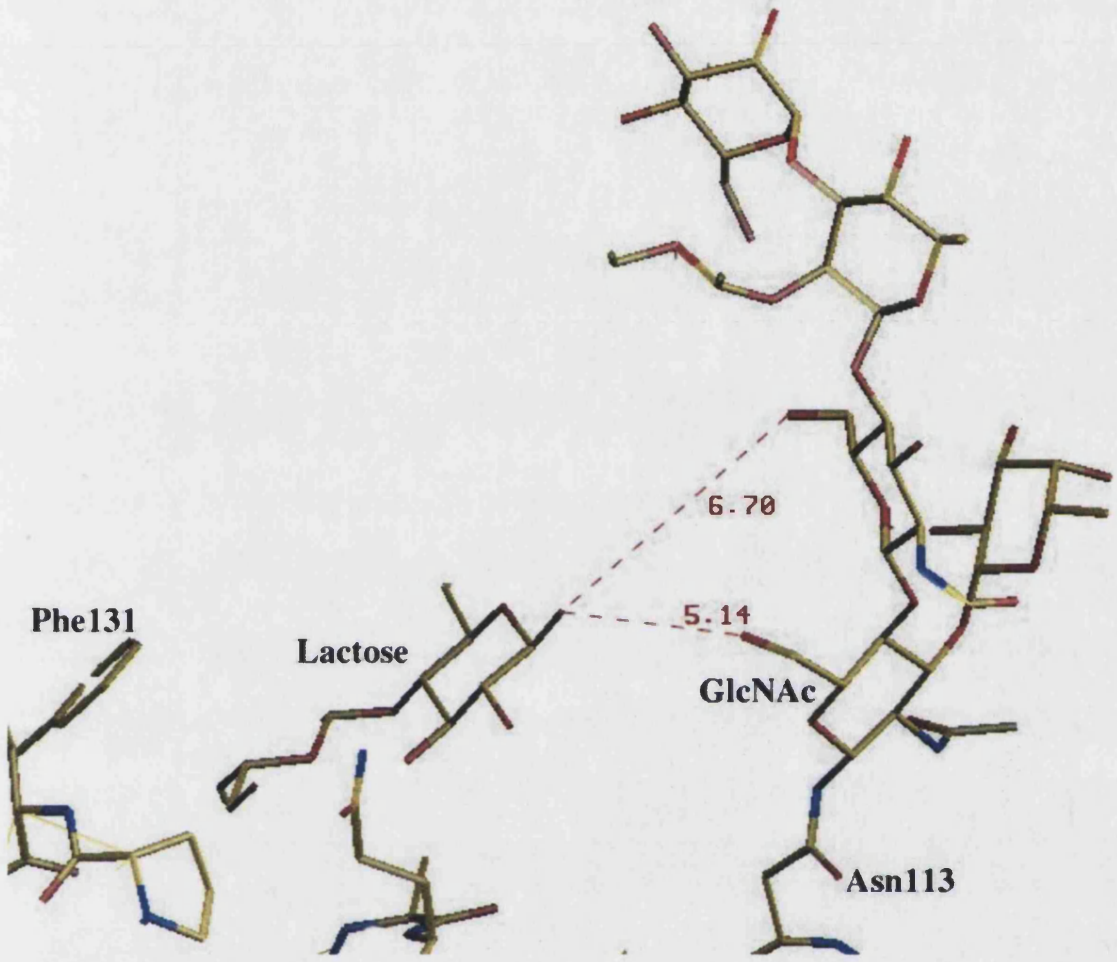
**Figure 23 - The Crystal Structures of Native and Recombinant ECL**




This figure shows the superposition of *rec*ECL-Lac (at 1.7Å resolution), in blue, on *n*ECL-Lac (at 2.0Å resolution), in pink. Three of the attached sugar units in the native structure are also shown in pink and the bound lactose molecules are shown in gold. The absence of the bound oligosaccharide from *rec*ECL does not seem to alter the overall tertiary structure of the lectin, as seen above. There are a few small differences in the positions of some loops but this probably results from the difference in resolution of the two crystal structures.

Comparison of the crystal structures of *n*ECL-Lac and *rec*ECL-Lac at 2.0Å and 1.7Å, respectively (Figure 23), revealed that there are no significant structural differences between native and recombinant forms of the lectin. This confirms that the presence of glycosylation does not influence the structure of the ECL protomer and has no effect on the structure of the combining site or carbohydrate binding. The first few sugar residues appear to be replaced by water molecules in the recombinant structure. Although small differences can be observed in the positions of some of the loop structures, these probably result from the difference in resolution between the two structures. There are no large structural rearrangements and the positions of the lactose and metal ions in the combining sites of the two ECL forms are equivalent.

**Figure 24 - Interaction Between the Lactose and Oligosaccharide**




This figure shows the long-range interactions (distances are in Ångstroms) between the N-linked oligosaccharide on molecule B in the asymmetric unit and the lactose molecule bound to a symmetry-related copy of molecule A. This is likely a result of the crystal packing, as it is not observed with molecule A and symmetry-related molecules.

However, possible communication was observed between the first sugar residue of the bound oligisaccharide in one protein molecule in the asymmetric unit (molecule B) with the glucose moiety of the lactose bound in the combining site of a symmetry-related molecule (of molecule A). The sugars are positioned approximately 5Å apart and oriented almost perpendicular to each other (Figure 24). As displayed in Figure 24, there are only two long-range interactions between the two sugar residues, between O1' of the lactose molecule and O6 of both GlcNAc molecules in the bound oligosaccharide. However, no communication was observed between the glycosylation on molecule A in the asymmetric unit and symmetry-related molecules. Instead, the symmetry-related molecules close to molecule A packed were oriented such that the

bound lactose molecules were positioned distant from the glycosylation on molecule A. This suggests that the "interaction" observed between molecule B and a symmetry-related copy of molecule A is merely a consequence of the crystal packing. Whilst it is crystallographically interesting, it does not represent an interaction that affects the biological activity of *n*ECL.

ECorL (Shaanan et al. 1991) and *Sven*ECL (Svensson et al. 2002) are reported to have two N-linked glycosylation sites at $Asn^{17}$ and $Asn^{113}$. In contrast, *n*ECL has only one glycosylation site – $Asn^{113}$. In the sequence of *n*ECL (Figure 13), the amino acid at position 17 is aspartic acid, not asparagine, and differs in this respect from ECorL and *Sven*ECL. The presence of a heptasaccharide bound to ECorL at $Asn^{17}$ was postulated to force the lectin into a non-canonical dimer structure and a similar rationale was applied to *Sven*ECL. However, the absence of bound carbohydrate at this position in *n*ECL indicates that there must be factors other than glycosylation influencing the quarternary association of legume lectin protomers. The dimeric structure of *n*ECL will be discussed in greater detail in section 3.3.3.4.

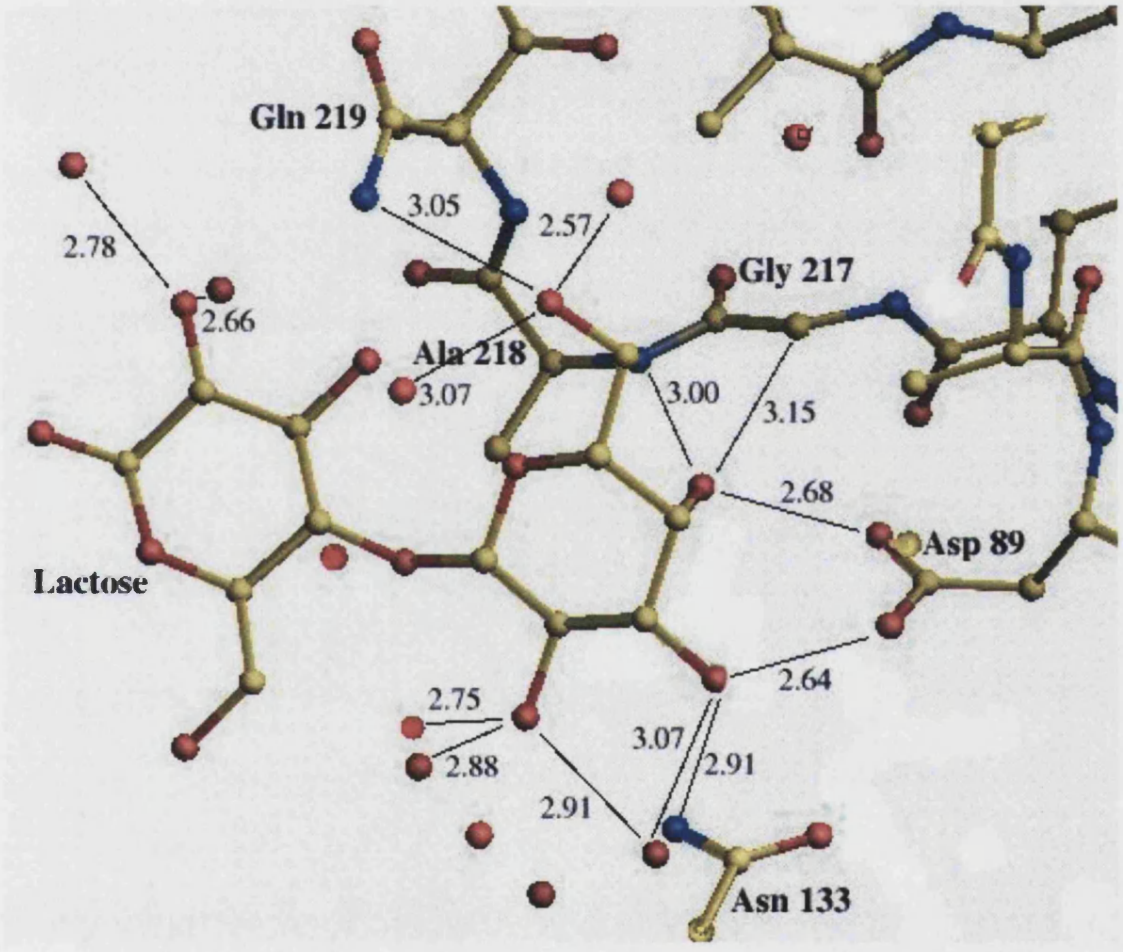
### 3.3.3.3 Lactose Binding

Analysis of contacts between lectin and lactose revealed that lactose binding by ECL is mediated by a set hydrogen bonds between the lectin and the disaccharide. Water molecules mediate indirect hydrogen bonds between $Leu^{86}$, $Gly^{107}$, $Asn^{133}$, $Ala^{218}$ and $Gln^{219}$ and the O2, O3 and O6 of galactose and O2 of glucose. In addition, Lac also makes contacts with further water molecules in the combining site. Some of the *rec*ECL-Lac contacts are displayed in Figure 25. This is similar to the mode of carbohydrate binding observed in ECorL, in which the galactose ring of lactose forms hydrophobic interactions with the side chains of $Ala^{88}$, $Tyr^{106}$, $Phe^{131}$ and $Ala^{218}$ and hydrogen bonds with the side chains of $Asp^{89}$ and $Asn^{133}$ (Shaanan et al. 1991). It has been suggested that the side chain of $Ala^{218}$ that prevents ECorL binding glucose or mannose by sterically hindering the optimal orientation of monosaccharides with an equatorial 4-OH in the combining site. Based upon similarities in the architecture of their combining sites and conservation of combining site residues, this would also hold true for ECL.

ECL has similar carbohydrate-binding properties to ECorL, although ECorL has a higher affinity for fucosyl-N-acetyllactosamine compared to LacNAc (Moreno et al. 1997; Teneberg et al. 1994). The altered carbohydrate specificities of *Sven*ECL and ECorL are postulated to be due to differences in their amino acid sequences at positions 111 and 125 (Svensson et al. 2002). The substitution of residues at these

positions causes rotation of the side chain of $Val^{92}$ and this is thought to induce structural changes in the combining site. Although there are no contacts between these three residues in the crystal structure of *n*ECL, $Val^{92}$ was found to make contacts with $Val^{126}$, which is located close to the combining site. Thus, it is possible that substitutions causing movement of $Val^{92}$ might affect carbohydrate binding, however, this remains to be proved.

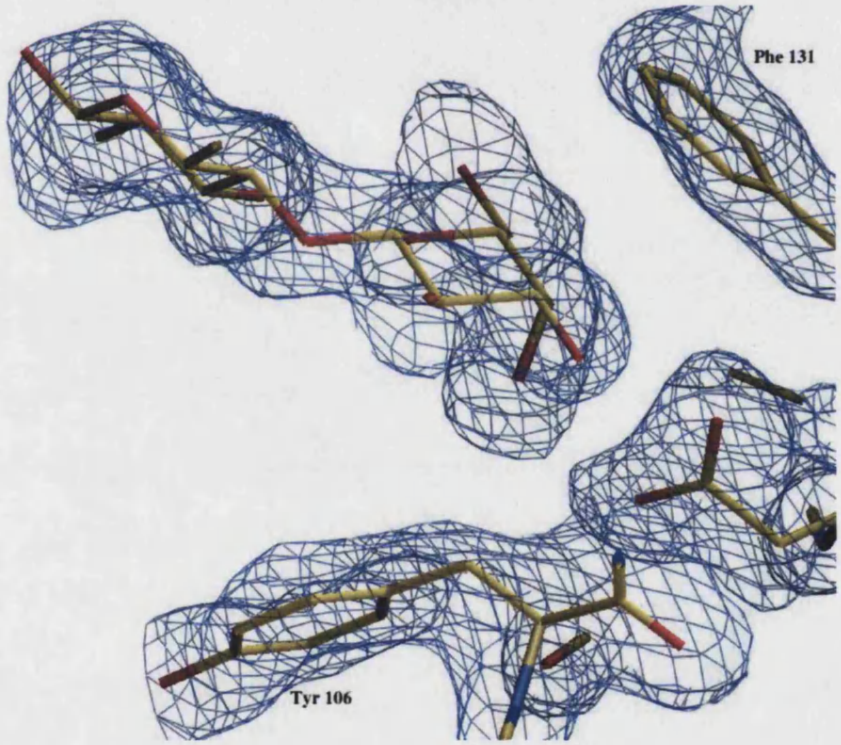
**Figure 25 - Interaction of Lactose with Combining Site Residues in *rec*ECL**




Lactose binding by ECL is mediated by a number of hydrogen bonds and van der Waals contacts with a number of amino acid side chains and water molecules in the combining site. Only the galactose moiety of the disaccharide is accommodated in the combining site in a manner analogous to that observed for ECorL. Lactose makes contacts with the side chains of $Asp^{89}$, $Gly^{107}$, $Asn^{133}$, $Ala^{218}$ and $Gln^{219}$ and carbonyl oxygen of $Leu^{86}$ and the galactose ring forms hydrophobic interactions with $Phe^{131}$. This figure displays several of the contacts formed between lactose and *rec*ECL – solid lines represent hydrogen bonds or van der Waals contacts, with all distances quoted in Ångstroms.

Hydrophobic stacking interactions were noted between the aromatic ring of Phe$^{131}$, the galactose ring of the bound lactose and the side chain of Tyr$^{106}$ (Figure 26). The side chain of Tyr$^{106}$ was found to adopt a generously allowed configuration in all of the ECL-Lac structures solved, as determined by the Ramachandran plot (Appendix 1). This is also true for *Sven*ECL (PDB code 1GZC) and the crystal structure of ECorL (PDB code 1AX0). However, in the unliganded *rec*ECL structure, only Tyr$^{106}$ in one of the four molecules (one of the three molecules with glycerol bound in the combining site) was located in an additional allowed region of the Ramachandran plot. Thus, it can be postulated that the configuration of this residue is affected by carbohydrate binding.

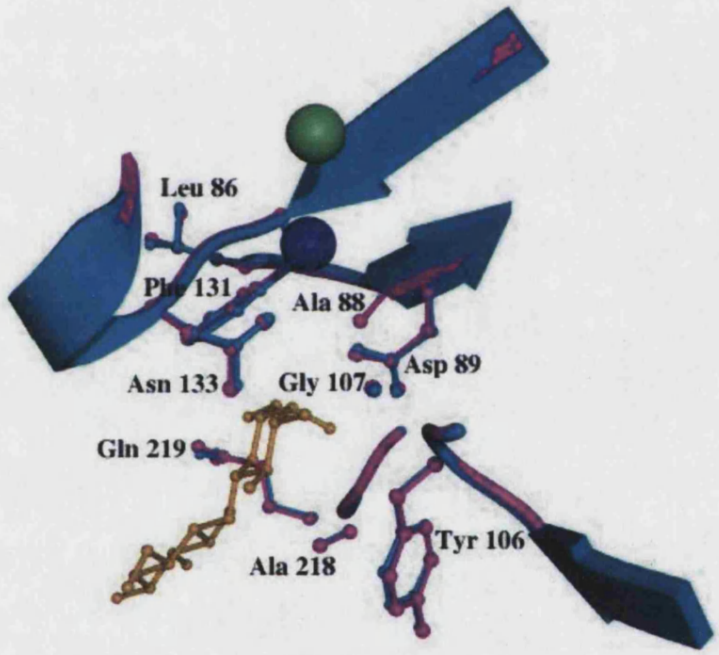
**Figure 26 - Hydrophobic Stacking Interactions Between Lactose and Combining Site Residues Tyr$^{106}$ and Phe$^{131}$**




The electron density map clearly outlines the hydrophobic stacking interaction between the galactose ring of the bound lactose and the side chains of combining site residues Tyr$^{106}$ and Phe$^{131}$. Interestingly, the side chain of Tyr$^{106}$ adopts a generously allowed configuration in the structures of the *n*ECL-Lac and *rec*ECL-Lac complexes, as was previously observed in the crystal structures of ECorL and ECL in complex with lactose (Shaanan et al. 1991; Svensson et al. 2002). However, the side chain of Tyr$^{106}$ in the *rec*ECL structure solved in the absence of carbohydrate did adopted allowed configurations. This suggests that the side chain of Tyr$^{106}$ is affected by carbohydrate binding.

The crystal structures of *n*ECL-Lac and *rec*ECL-Lac complexes at 2.0Å and 1.7Å, respectively, superimposed with an overall rmsd of 0.29Å. Comparison of the spatial arrangement of residues in the combining sites of both forms of the lectin revealed that *rec*ECL binds lactose in the same way as *n*ECL. Further to the observed equivalence of carbohydrate binding by *n*ECL and *rec*ECL (Stancombe et al. 2003), this confirms that *rec*ECL is native-like terms of its carbohydrate-binding activity as well as its tertiary structure. This analysis also confirms that glycosylation of ECL does not influence its carbohydrate binding properties or the structure of the combining site.

**Figure 27 - Comparison of the Combining Sites of *rec*ECL in Lactose-Bound and Unliganded Forms**




This figure (prepared using Molscript, Povray and GIMP) displays a superposition of combining site residues in the crystal structures of *rec*ECL in lactose-bound and unliganded forms. The structure of *rec*ECL in the lactose-bound conformation is shown in blue with $Mn^{2+}$ and $Ca^{2+}$ depicted as green and navy spheres, respectively and the bound lactose molecule in gold. The structure of *rec*ECL with no ligand bound is shown in pink. The comparison clearly shows that there are no significant structural changes, with respect to the residues in the combining site, upon lactose binding. Instead, the amino acid side chains that are involved in carbohydrate recognition appear to be optimally spatially oriented in the ECL protomer, even in the absence of bound ligand.

The structures of *rec*ECL unliganded and in complex with Lac (at 2.13Å and 1.70Å resolution, respectively) superimposed with an overall rmsd of 0.21Å. A direct comparison of side chain positions in the combining sites of the two structures (Figure

27) revealed no structural rearrangements within the combining site upon lactose binding. This indicates that the amino acids involved in carbohydrate recognition are optimally oriented in the ECL protomer. The calcium and manganese ions maintain the orientation of key carbohydrate-binding residues within the combining site and the conserved *cis*-peptide bond correctly positions the side chain of $Asp^{89}$. Analysis of the crystal structure of ECorL indicates that residues in the combining site are optimally oriented for carbohydrate binding in this lectin, even in the absence of ligand (Shaanan et al. 1991). So, it appears that lectins recognise their ligands via a "lock-and-key" mechanism unlike enzymes, which exhibit an "induced-fit" mode of substrate binding. This helps to explain the specificity of lectins for monosaccharides, since rigidity in the combining site would prevent galactose-specific lectins, for example, altering their shape to accommodate glucose or fucose even though these sugars are epimers of galactose.

#### 3.3.3.4 Mode of Dimerisation

ECL is biologically active as a dimer and it is not possible to isolate the lectin in monomeric form (J.A. Chaddock, personal communication). Protomers of *n*ECL and *rec*ECL associate back-to-back, forming dimers via the "handshake motif" in which the two protomers are tilted with respect to each other (Figure 28). As previously mentioned, this non-canonical mode of dimerisation was originally observed for ECorL (Shaanan et al. 1991) and was also recently reported for *Sven*ECL (Svensson et al. 2002). It was proposed that ECorL and *Sven*ECL dimerise in this manner because the presence of a heptasaccharide N-linked to $Asn^{17}$ prevents the formation of a canonical dimer. However, this glycosylation site is absent from the *n*ECL studied in this investigation and *rec*ECL is not glycosylated because it is recombinantly expressed in *E.coli*, yet protomers of both *n*ECL and *rec*ECL associate in the same way (i.e. back-to-back). This observation throws doubt on the proposed influence of glycosylation on the quarternary association of legume lectins and points to the potential importance of small sequence changes in determining protein structure (as discussed in (Manoj and Suguna 2001).

Examination of legume lectin sequences, focusing on the regions forming interfaces between protomers and dimers, identified amino acid residues that might influence the mode of oligomerisation (Manoj and Suguna 2001). Extrapolation of these results to ECL reveals that the primary structure of this lectin (Figure 13) contains features that indicate it could be expected to form ECorL-type (i.e. X3-type) dimers. Legume lectins that do not form canonical dimers have charged residues at two critical

positions, equivalent to $Phe^{11}$ and $Lys^{55}$ in ECL. Furthermore, the presence of $Arg^{73}$ and $Lys^{171}$ in ECL indicates the formation of a non-canonical dimer, as arginine and lysine residues at equivalent positions are unique to those legume lectins that dimerise via the handshake motif. Although this sequence analysis would predict the mode of dimerisation observed for ECL, it should not be considered definitive and cannot be used to accurately predict the quarternary structures of legume lectins that have not been structurally characterised (Manoj and Suguna 2001).

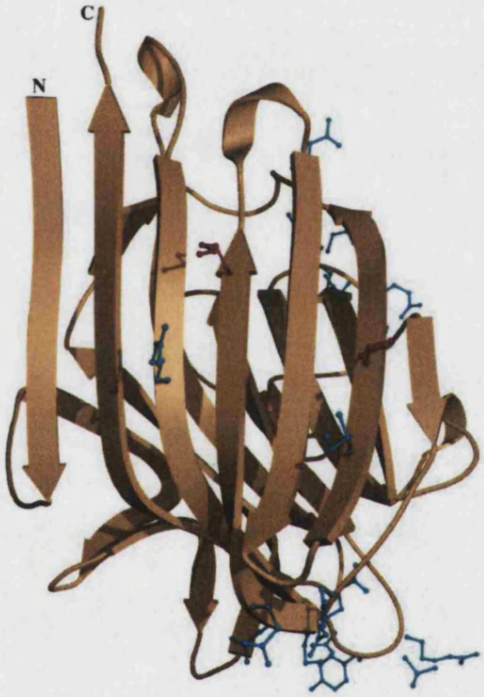
**Figure 28 - Quarternary Structure of ECL**

Native and recombinant forms of ECL adopt the same quarternary structure, forming ECorL-like dimers in which two protomers associate back-to-back via the "handshake motif". The dimer interface is stabilised by residues located in the flat, six-stranded β-sheet and loops connecting this sheet to the curved, seven-stranded β-sheet (Figure 29). This figure displays the dimeric structure of nECL in complex with lactose (in pink) and shows that the combining site is located on the opposite face of the lectin to the dimer interface. The calcium and manganese ions are shown as red and blue spheres, respectively. The bound oligosaccharide (not shown) is positioned on the curved seven-stranded β-sheet, distant from the dimer interface, and does not influence the quarternary structure of ECL. This figure was prepared using Molscript, Povray and GIMP.

The dimer interface is stabilised by two hydrogen bonds, between $Lys^{171}$ in one protomer and $Thr^{193}$ in the other, and several van der Waals contacts involving

residues $Arg^{73}$, $Glu^{79}$, $Gln^{80}$, $Pro^{81}$, $Tyr^{82}$, $Thr^{83}$, $Arg^{84}$, $Lys^{116}$, $Gln^{117}$, $Asp^{118}$, $Asn^{119}$, $Asn^{148}$, $Asp^{161}$, $Asn^{162}$, $Gln^{164}$, $Lys^{171}$, $Ile^{191}$, $Thr^{193}$, $Gln^{202}$, $Val^{203}$ and $Asp^{221}$. The amino acids mediating protomer-protomer contacts are located in four strands of the six-stranded, anti-parallel β-sheet and in loops connecting this sheet to the curved, seven-stranded β-sheet (Figure 29). The N- and C-termini associate together, forming the first two strands of the six-membered sheet, but play no role in dimerisation of the lectin because the protomers are tilted with respect to each other such that there are no contacts between the N- and C-termini of different protomers. Thus, only the remaining four strands of the six-membered sheet associate together in the dimer structure.

**Figure 29 - Amino Acid Residues That Stabilise the Dimer Interface in ECL**




The structure of the ECL monomer is oriented such that the flat, six-stranded β-sheet is at the forefront. The N- and C- termini are labelled and form the first two strands of the anti-parallel sheet. The amino acid residues that are involved in forming hydrogen bonds that stabilise the dimer interface ($Lys^{171}$ and $Thr^{193}$) are coloured in red. Residues that stabilise the dimer interface by forming hydrophobic or van der Waals contacts are coloured in blue. Most of these residues are located in the third to sixth strands of the sheet and in loops connecting the six-stranded sheet to the curved seven-stranded β-sheet. This figure was prepared using the programs Molscript, Povray and GIMP.

### 3.3.4 Comparison with the Previously Reported Structure

As previously mentioned, the crystal structures of natively-sourced ECL in complex with 2-fucosyllactose and lactose were recently reported (Svensson et al. 2002). *Sven*ECL protomers adopt the legume lectin fold and associate via the handshake motif. However, several differences have been noted between *Sven*ECL and the crystal structures of *n*ECL and *rec*ECL determined in this investigation.

**Figure 30 - Analysis of the Primary Structure of *rec*ECL**

```
recECL      VETISFSFSE FEPGNNDLTL QGAAIITQSG VLQLTKINQN
SvenECL     VETISFSFSE FEPGNDNLTL QGAALITQSG VLQLTKINQN

recECL      GMPAWDSTGR TLYTKPVHIW DMTTGTVASF ETRFSFSIEQ
SvenECL     GMPAWDSTGR TLYTKPVHMW DSTTGTVASF ETRFSFSIEQ

recECL      PYTRPLPADG LVFFMGPTKS KPAQGYGYLG VFNNSKQDNS
SvenECL     PYTRPLPADG LVFFMGPTKS KPAQGYGYLG VFNNSKQDNS

recECL      YQTLAVEFDT FSNPWDPPQV PHIGIDVNSI RSIKTQPFQL
SvenECL     YQTLAVEFDT FSNPWDPPQV PHIGIDVNSI RSIKTQPFQL

recECL      DNGQVANVVI KYDASSKILL AVLVYPSSGA IYTIAEIVDV
SvenECL     DNGQVANVVI KYDAPSKILH VVLVYPSSGA IYTIAEIVDV

recECL      KQVLPEWVDV GLSGATGAQR DAAETHDVYS WSFHASLPET N
SvenECL     KQVLPDWVDV GLSGATGAQR DAAETHDVYS WSFQASLPE- -
```

The primary structure of *n*ECL differs at ten positions from the sequence of ECL reported by Svensson et al, 2002 (*Sven*ECL). The ten differences are highlighted in green and the N-linked glycosylation sites are underlined. None of the differences affect residues involved in carbohydrate binding (highlighted in red) or the dimer interface (in blue).

In total, there are ten differences between the amino acid sequence of *n*ECL and *Sven*ECL (Figure 30). These are: $N^{16}D$, $D^{17}N$, $I^{25}L$, $I^{59}M$, $M^{62}S$, $S^{175}P$, $L^{180}H$, $A^{181}V$, $E^{206}D$ and $H^{234}Q$ (where the first residue is that in *n*ECL and the second is that in *Sven*ECL). The differences at positions 25, 181 and 206 represent conservative substitutions, where the chemical nature of the residue is not changed. The residues at positions 16 and 17 appear to be swapped in *n*ECL compared with *Sven*ECL and although this alters the chemical nature of the residue each position, the side chains

are of the same size and shape and cannot be distinguished from one another on the basis of the electron density pattern. The remaining variations are not conservative substitutions.

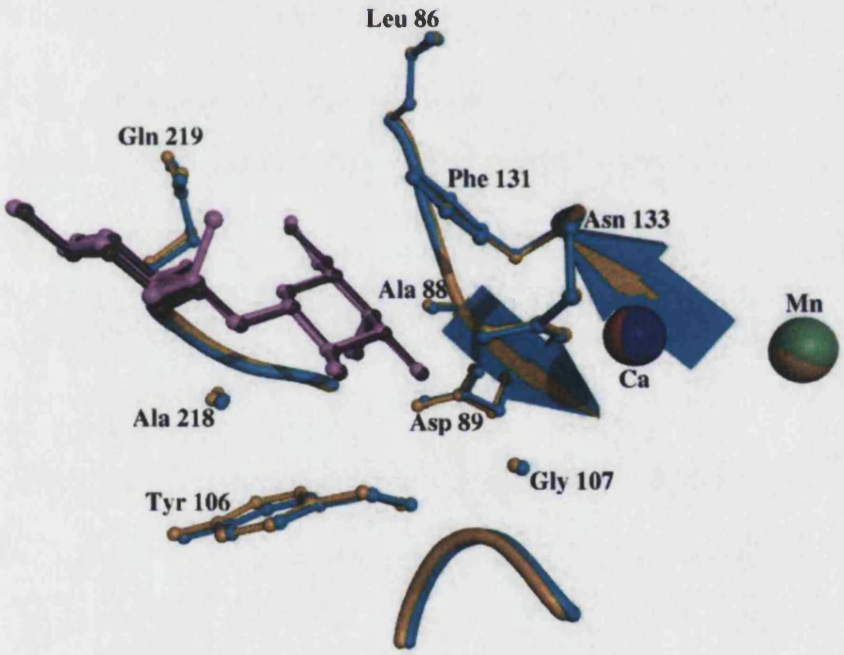
Given that the sequence of *Sven*ECL determined by peptide mapping (briefly described in section 3.1.5.3) differs from the sequence indicated by the electron density map of the protein (Svensson et al 2002), it does not appear to be a reliable method for determining protein primary structure. The amino acid sequence of *rec*ECL, on the other hand, was determined by DNA sequencing using material from two independent sources of *E.cristagalli* (Stancombe et al. 2003) and careful analysis of the calculated electron density maps confirmed this sequence at all ten of the locations where it differs from *Sven*ECL.

Differences between the sequences of *rec*ECL and *Sven*ECL might represent subtle variations across the species.None of the sequence differences between *Sven*ECL and *n*ECL affect amino acids involved in carbohydrate binding or stabilising the dimer interface. However, a potential N-linked glycosylation site at residue 17 is absent from *n*ECL. *Sven*ECL is glycosylated at two positions ($Asn^{17}$ and $Asn^{113}$), the same as ECorL, but although weak electron density for the sugars was observed, the bound oligosaccharides were not modelled in the crystal structures (Svensson et al. 2002). *Sven*ECL crystallised under different conditions and in a different space group to both *n*ECL and *rec*ECL. As described in section 3.3.1, *n*ECL co-crystallised with Lac in space group $P6_5$ and *rec*ECL crystallised in three forms (P1, $P2_1$, and $P3_212$) and co-crystallised with Lac in space groups $P2_1$ and C2. *Sven*ECL was co-crystallised with lactose and 2'-fucosyllactose in space group $P4_32_12$ (Svensson et al. 2002). The structure of the *Sven*ECL-lactose complex at 1.6Å resolution (PDB code 1GZC) superimposed onto the structure of the *n*ECL-Lac complex (at 2.0Å resolution) with an overall rmsd of 0.32Å.

The equivalence of *n*ECL and *Sven*ECL in terms of their tertiary structures is not surprising, considering the conservation of structure among legume lectins. These two forms of ECL are 96% identical in terms their primary structures. Sequence identity among pairs of legume lectins that have been structurally characterised varies from 28 to 99% (Manoj and Suguna 2001) yet they all adopt the same tertiary structure characterised by two anti-parallel β-sheets (Figure 11). *n*ECL shares 96% sequence identity with ECorL, which was used as a model for phase determination by molecular replacement. This high level of sequence identity indicates that the proteins are likely to fold in a similar manner. Therefore, the tertiary structures of *Sven*ECL and *n*ECL would be expected to be almost identical.

The combining site of *n*ECL was compared with that of *Sven*ECL in complex with lactose (Figure 31). There are no significant differences between the positions of the amino acid residues involved in carbohydrate binding in the two structures. In both crystal structures, the side chain of $Tyr^{106}$ was found to adopt a generously allowed configuration (according to the Ramachandran plot). In conjunction with $Phe^{131}$, $Tyr^{106}$ makes a hydrophobic stacking interaction with the galactose ring of the bound lactose. None of the sequence differences between *n*ECL and *Sven*ECL alter the amino acid composition of the combining site, so both forms of the lectin would be predicted to bind carbohydrates in the same way.

**Figure 31 - Comparison of Combining Sites of *n*ECL and the Previously Reported Structure of ECL in Complex with Lactose**




The combining site of *n*ECL is shown in blue, with the manganese and calcium ions in green and navy, respectively, and lactose in purple. The previously reported structure of ECL (Svensson et al. 2002; PDB code 1GZC) is shown in gold, with manganese and calcium in orange and red, respectively, and lactose in pink. This diagram (produced using Molscript, Povray and GIMP) shows that there are no significant differences in the positions of the combining site residues between the two structures.

*Sven*ECL forms ECorL-like (i.e. back-to-back) dimers as do both *n*ECL and *rec*ECL. The presence of bound carbohydrate at $Asn^{17}$ in *Sven*ECL was proposed to prevent the formation of a canonical dimer (as was originally postulated for ECorL). However, this residue is not glycosylated in the structure of *n*ECL, suggesting that

there are other factors intrinsic to the protomers that influence the quarternary structure of the lectin. This hypothesis is further supported by the observation that *rec*ECL protomers, which are devoid of glycosylation, also associate via the handshake motif.

## 3.4 SUMMARY OF RESULTS

Native and recombinant forms of ECL were crystallised in five space groups ($P6_5$, P1, $P2_1$, C2 and $P3_212$) under a variety of conditions. Whilst *n*ECL could only be crystallised in complex with Lac, *rec*ECL was crystallised in the presence and absence of carbohydrate. A total of six crystal forms of ECL were analysed and the structures were solved using molecular replacement with ECorL as a search model. All residues except the C-terminal $Asn^{241}$ could be modelled in the observed electron density.

Protomers of *n*ECL and *rec*ECL adopt the conserved legume lectin fold that is characterised by two anti-parallel β-sheets and the presence of $Ca^{2+}$ and $Mn^{2+}$ close to the carbohydrate-combining site. The overall fold comprises one flat, six-stranded β-sheet; one curved, seven-stranded β-sheet and; one short five-membered β-sheet and these structural elements are linked by a series of loops. The combining site is located on the surface of the seven-stranded β-sheet, formed by four loops associated with this face of the lectin. The N- and C-termini associate together forming the first two strands of the six-stranded β-sheet. Protomers of both forms of ECL associate back-to-back, forming dimers via the "handshake" motif, as was observed for ECorL. The dimer structure is stabilised by two hydrogen bonds and several van der Waals contacts between residues in four strands of the flat, six-stranded β-sheet and in loops associated with these structural elements.

Equivalence of carbohydrate specificity in *n*ECL and *rec*ECL has already been documented (Stancombe et al. 2003). Direct comparison of the structure of the combining site in the *n*ECL-Lac complex with that in the *rec*ECL-Lac complex revealed no differences in either the spatial arrangement of combining site residues or in the orientation of the bound disaccharide. Furthermore, there are no significant differences in the overall fold of the *rec*ECL protomer compared to *n*ECL or in the structures of the dimers they form. These results clearly demonstrate that *rec*ECL is native-like in terms of both its structure and mode of carbohydrate binding. The only difference between native and recombinant forms of ECL is the absence of bound carbohydrate from *rec*ECL. The amino acid sequence of *n*ECL contains one potential N-linked glycosylation site, $Asn^{113}$, to which a heptasaccharide is N-linked in the mature protein. The structure of the oligosaccharide bound to *n*ECL was modelled into the observed electron density and was not found to make contacts with any part of the lectin except

the asparagine to which it is linked. The role of carbohydrate on lectins is unknown and, in the case of ECL, the bound heptasaccharide does not appear to affect carbohydrate recognition or influence quarternary structure.

The structure of the combining site in ECL in the absence of carbohydrate is not significantly different to that in lactose-bound forms of the lectin. This indicates that key carbohydrate-binding residues are optimally positioned in combining site of the ECL protomer and suggests that lectins recognise their ligands via a "lock and key" mechanism. ECL binds lactose by forming a network of indirect hydrogen bonds with the galactose moiety through a number of structural water molecules. The side chains of $Phe^{131}$ and $Tyr^{106}$ form hydrophobic stacking interactions with the pyranose ring of galactose and there are several van der Waals contacts between lectin and carbohydrate. Key residues implicated in lactose binding were identified as $Leu^{86}$, $Asp^{89}$, $Tyr^{106}$, $Gly^{107}$, $Phe^{131}$, $Asn^{133}$, $Ala^{218}$ and $Gln^{219}$.

A different crystal form of native ECL (space group $P4_32_12$) was reported in 2002 that differs in primary structure from *n*ECL and *rec*ECL in ten positions. Only one of the amino acid substitutions alters the structure of the lectin by removing a potential N-linked glycosylation site at position 17. In the reported structure (Svensson et al. 2002), $Asn^{17}$ is glycosylated and electron density for the sugar was observed at this position, although no sugars were modelled. Despite the presence of carbohydrate at $Asn^{17}$, the tertiary and quarternary structures reported for native ECL do not differ significantly from those determined for *n*ECL and *rec*ECL in this investigation.

# CHAPTER FOUR
# STRUCTURAL STUDIES ON NON-TOXIC FRAGMENTS OF BOTULINUM NEUROTOXIN A

## 4.1 INTRODUCTION

The botulinum neurotoxins (BoNTs) are among the most poisonous proteins known to Man and are the causative agents of the disease botulism. The seven BoNT serotypes (A, B, C1, D, E, F and G) are produced and secreted by strains of the anaerobic bacteria *Clostridium botulinum*, *Clostridium barati* and *Clostridium butyricum*. These toxins cause descending flaccid paralysis by disrupting neurotransmission at the neuromuscular junction. Together with tetanus neurotoxin (TeNT), the BoNTs make up the Clostridial Neurotoxin (CNT) family. TeNT is produced and secreted by *Clostridium tetani* and causes the disease tetanus, characterised by spastic paralysis. The mouse $LD_{50}$ for the clostridial neurotoxins is 0.1-1.0 ng/kg (Schiavo et al. 2000) and, if medical attention is not sought rapidly, both tetanus and botulism can be fatal. However, despite their potent toxicity, clostridial neurotoxins have become useful tools in medical research and are used as pharmaceutical therapeutic agents for a variety of conditions.

**Nota bene:** A review of recent research into the structures, functions and therapeutic applications of clostridial neurotoxins (Turton et al. 2002) is presented in Appendix 3.

### 4.1.1 Botulism and Tetanus

#### 4.1.1.1 Botulism

Although botulism was first described in the eighteenth century, the bacteria and proteins responsible were not identified until the 1890s. The BoNT serotypes associated with botulism in humans are A, B and E, with few cases reportedly caused by BoNT/F (Montecucco et al. 1996). There are three recognised types of botulism in humans: food-borne, wound and infant botulism. The most common form is food-borne botulism, an intoxication resulting from the ingestion of foods contaminated with *C.botulinum* spores that have germinated and produced BoNT. Wound botulism is comparatively rare and is caused by the germination of *C.botulinum* spores in open wounds. Symptoms of both types normally appear within 12-36 hours of intoxication or infection. The disease proceeds with symptoms of food poisoning (in the case of food-borne botulism), dry mouth and visual impairment, followed by general muscle weakness and a rapidly progressing, descending flaccid paralysis. If left untreated,

botulism is fatal because respiratory failure or cardiac arrest will occur (Niemann 1991). Infant botulism is a toxic infection that results from germination and proliferation of *C.botulinum* spores in the neonate intestine in the absence of competing bacterial flora (Montecucco et al. 1996). Symptoms differ slightly from those reported for food-borne and wound botulism and include lethargy, listlessness, hypotonia, feeble cry, poor feeding, weakness, drooping eyelids and apnoea.

At present there is no cure for botulism, but few cases of the disease are fatal due to advances in medical care. Early diagnosis is important and treatment involves the administration of antitoxin, comprising antibodies raised against BoNTs A, B and E (Robinson and Nahata 2003). Respiratory monitoring is essential because even early administration of antitoxin cannot prevent paralysis and respiratory arrest may occur. Recovery is slow and depends upon the BoNT serotype involved. Different serotypes are active for differing periods of time, with BoNT/A exhibiting the longest duration of action (paralysis lasts for up to 4-6 months). The differences in duration of action are due to replenishment of substrate proteins (observed for BoNTs E and F) or longevity of protease action, as for BoNTs A, B and C1 (Foran et al. 2002). The fact that full recovery is possible indicates that BoNTs do not cause lasting structural damage to neurons, although complete restoration of exocytosis and relief from paralysis cannot occur until the toxin has been fully degraded (Ahnert-Hilger and Bigalke 1995). Nerve sprouting from neurons poisoned by BoNT/A creates new synapses that mediate neurotransmission, allowing the onset of recovery (Meunier et al. 2003). These neurite sprouts are retracted upon restoration of activity to poisoned synapses (dePaiva et al. 1999). Thus, full restoration of synaptic function is possible without the need for permanent synaptic remodelling.

There is no widely available anti-botulism vaccine, but a pentavalent toxoid vaccine that protects against serotypes A-E is administered to people deemed "at risk" of infection, for example health workers and members of the armed forces. The vaccine is manufactured by detoxifying BoNTs A-E with formalin, but it is expensive to produce because it requires dedicated facilities and can produce some unpleasant side effects due to the presence of residual formaldehyde. Efforts are underway to produce recombinant vaccines, as this would remove the need for a dedicated manufacturing facility, reduce production costs and eliminate the side effects caused by formaldehyde. Fragments of BoNTs A, C and F have been recombinantly expressed and used to induce immune responses in mice (Byrne et al. 1998; Byrne et al. 2000; Clayton et al. 1995; Kiyatkin et al. 1997; LaPenotiere et al. 1995). The potential of DNA vaccination

has also been investigated and it is thought to be a viable method of protection (Clayton and Middlebrook 2000).

#### 4.1.1.2 Tetanus

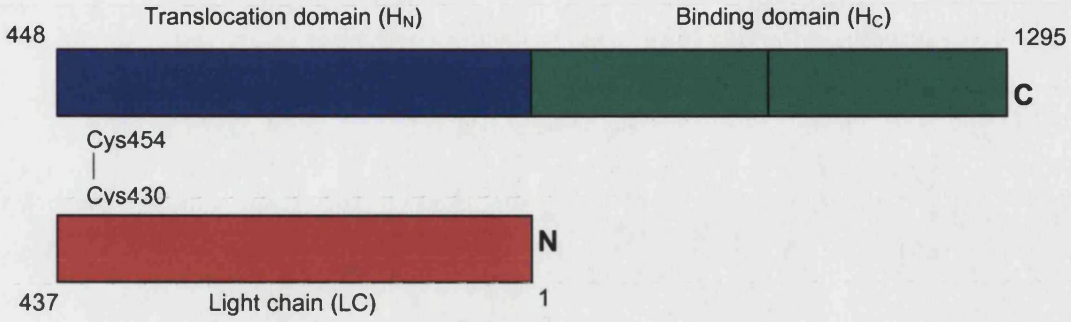
Cases and symptoms of tetanus have long been known, but the bacterium responsible for the disease was not isolated until 1891 (Schiavo et al. 2000). Spores of *C.tetani* germinate under anaerobic conditions, such as those in a deep wound, producing TeNT. Initial symptoms include headache, neck and back pains, fever and muscle stiffness, which proceeds to an overall spastic paralysis with generalised reflex spasms (Niemann 1991). Respiratory failure occurs if medical attention is not sought quickly. The opposite symptoms of tetanus and botulism result from the different sites of action of TeNT and BoNTs, although at high concentrations TeNT may induce symptoms of botulism (Singh et al. 1995). Treatment of tetanus involves the use of muscle relaxants and artificial respiration and full recovery is possible with no lasting nervous damage. An effective toxoid vaccine, composed of formaldehyde-treated TeNT, is widely available and its use has almost eliminated tetanus from the developed world (Niemann 1991).

### 4.1.2 Molecular Biology of Clostridial Neurotoxins

CNTs are synthesised in the bacterial cytosol as single-chain polypeptides. These precursor polypeptides are post-translationally nicked within a surface-exposed loop to form di-chain molecules of ~150kDa in which the light and heavy chains are linked by a single disulphide bond. This proteolytic nicking activates the toxin, as has been observed for several other bacterial toxins, including *E.coli* labile toxin, cholera toxin and diphtheria toxin (Lencer et al. 1997; Tsuneoka et al. 1993). The proteases responsible for CNT activation remain to be identified and might be serotype-specific because BoNTs B and E are not substrates for the protease that nicks BoNT/A (Dekleva and DasGupta 1989).

CNTs comprise three functional domains (Figure 32) each of which is involved in the mechanism of action. The ~50kDa light chain (LC) is the catalytic domain with zinc-dependent protease activity. The toxin heavy chain contains two domains, each of ~50kDa. The N-terminal half ($H_N$) is the translocation domain, which is known to form ion-conducting channels, and the C-terminal half ($H_C$) is the binding domain, which targets the toxins to cholinergic and CNS neurons.

**Figure 32 - Cartoon Depicting the Di-Chain Structure of Botulinum Neurotoxin A**




Botulinum neurotoxin A, a typical clostridial neurotoxin, is synthesised as a single-chain polypeptide and is post-translationally nicked to form an active di-chain molecule in which the light and heavy chains are linked by a disulphide bond ($Cys^{430}$-$Cys^{454}$). Within their di-chain structures, the clostridial neurotoxins are divided into three functional domains, each of approximately 50kDa. The light chain (depicted in red) is the catalytic domain, which acts as a zinc-dependent endopeptidase. The translocation domain (in blue) and the binding domain (in green) make up the heavy chain. The translocation domain forms ion channels in lipid bilayers and the binding domain facilitates internalisation of the toxins into their target neurons.

The eight CNTs are thought to have developed from a single ancestral gene because they exhibit 30-40% amino acid sequence similarity among their light and heavy chains (Ahnert-Hilger and Bigalke 1995). TeNT shares ~65% sequence homology (and ~35% identity) with the BoNTs (Lacy and Stevens 1999) and sequence alignment of the toxins (Appendix 4) reveals areas of high similarity. Further evidence for the existence of a common ancestor comes from the observation that neurotoxin gene elements are mobile and that gene transfer may take place between toxigenic and non-toxigenic strains of *C.botulinum* via phages, plasmids or conjugation transposons (Schiavo et al. 2000). Indeed, strains of the bacterium have been isolated that produce a "mosaic" BoNT, containing a mixture of elements from serotypes C and D (Moriishi et al. 1996a; Moriishi et al. 1996b).

CNTs are substrates for protein tyrosine kinases (Ferrer-Montiel et al. 1996) and phosphorylation increases their catalytic activity and thermal stability. The catalytic activity of BoNTs A and E increases with increased tyrosine phosphorylation and it has been postulated that CNTs are biologically active only in the phosphorylated form. Phosphorylation is accompanied by conformational changes that make the protein structure more compact, which may explain the enhanced catalytic activity (Encinar et al. 1998). The light and heavy chains of TeNT and BoNTs A, B, E are phosphorylated by Src, but serine/threonine kinases do not appear to act upon CNTs (Ferrer-Montiel et

al. 1996). This observation suggests that cell-signalling pathways in the neuron might be involved in the mechanism of action of the clostridial neurotoxins.

#### 4.1.2.1 Neurotoxin-Associated Proteins

BoNT/A has been purified as a 900kDa complex, comprising the 150kDa neurotoxin and a non-toxic haemagglutinin component, and it is this complex that is the causative agent of botulism (Chen et al. 1998). The complex has 360-fold higher oral toxicity in rats (Ohishi et al. 1977) and higher enzymatic activity (Cai et al. 1999) than the neurotoxin component alone. The non-toxic component of the complex consists of several proteins (known as neurotoxin-associated proteins or NAPs), some of which have haemagglutinating activity. The NAPs appear to be translated from a polycistronic mRNA, indicating that the neurotoxin and NAP genes are expressed simultaneously (Singh et al. 1995). In addition to the 900kDa complex, two smaller BoNT/A complexes of 500 and 300kDa have been purified, although the 300kDa complex does not exhibit haemagglutinin activity (Sagane et al. 2000).

Following ingestion, BoNTs are taken up by epithelial cells. The mouth is the least effective site of absorption, with the stomach intermediate and the intestine the most effective (Maksymowych and Simpson 1998). The 900kDa complex withstands the acidic, protease-rich conditions of the digestive tract but dissociates in the alkaline environment of the intestine. The haemagglutinins aid absorption of the complex to intestinal epithelial cells (Fujinaga et al. 1997). The neurotoxin is transcytosed to the mucosal side of the epithelium and enters the lymph and circulatory systems.

TeNT, on the other hand, is not released in complex with NAPs. It has never been isolated as a complex and there is no open reading frame upstream of the TeNT gene (Singh et al. 1995). Considering the protective role of NAPs in oral BoNT intoxication, the absence of NAP genes from the *C.tetani* genome may explain why TeNT is not a food poison.

#### 4.1.2.2 Do CNTs Form a Novel Class of Zinc Protease?

CNTs have been described as a new class of zinc protease because they exhibit several properties that have not been observed for other zinc-dependent endopeptidases (Montecucco and Schiavo 1993): Firstly, CNTs are synthesised as inactive protein precursors that require specific proteolytic cleavage for activation (Pellizzari et al. 1999). Whilst this has been observed for other bacterial toxins, it is not known for other zinc-dependent proteases (Montecucco and Schiavo 1993). CNTs are also inactive outside their target cells (whereas many zinc proteases remain active in extracellular environments). In addition, CNTs demonstrate extremely high substrate

specificity and show little sequence similarity with other zinc proteases in the region of the zinc-binding motif (Montecucco and Schiavo 1993). There are two types of zinc-dependent endopeptidase, based upon the mode of zinc coordination, astacin-like and thermolysin-like. However, whilst the CNT LCs share greater sequence and structural similarity with thermolysin, coordination of $Zn^{2+}$ in CNTs involves five ligands, as in astacin. Removal of $Zn^{2+}$ from BoNT abolishes its catalytic activity, but the substitution of other divalent metal ions for $Zn^{2+}$ in TeNT supports proteolysis, a feature not observed for other zinc proteases (Tonello et al. 1997).

It has been postulated that, unlike in other zinc proteases, the $Zn^{2+}$ plays both a structural and a functional role in CNTs, although this issue remains unresolved. Structural studies following the removal and replenishment of $Zn^{2+}$ showed that the activity of BoNT/A cannot be recovered, suggesting that loss of the zinc ion results in irreversible structural changes (Fu et al. 1998). However, this was contradicted by the observation that $Zn^{2+}$-depleted BoNT/A can block neurotransmission by binding zinc in the neuron cytosol, indicating that loss of $Zn^{2+}$ does not lead to irreversible structural changes (Simpson et al. 2001). Further studies into the proposed structural role of zinc in BoNTs (Li and Singh 2000a) demonstrated partial recovery of catalytic activity following removal and addition of $Zn^{2+}$, accompanied by some irreversible structural changes that do not block zinc binding.

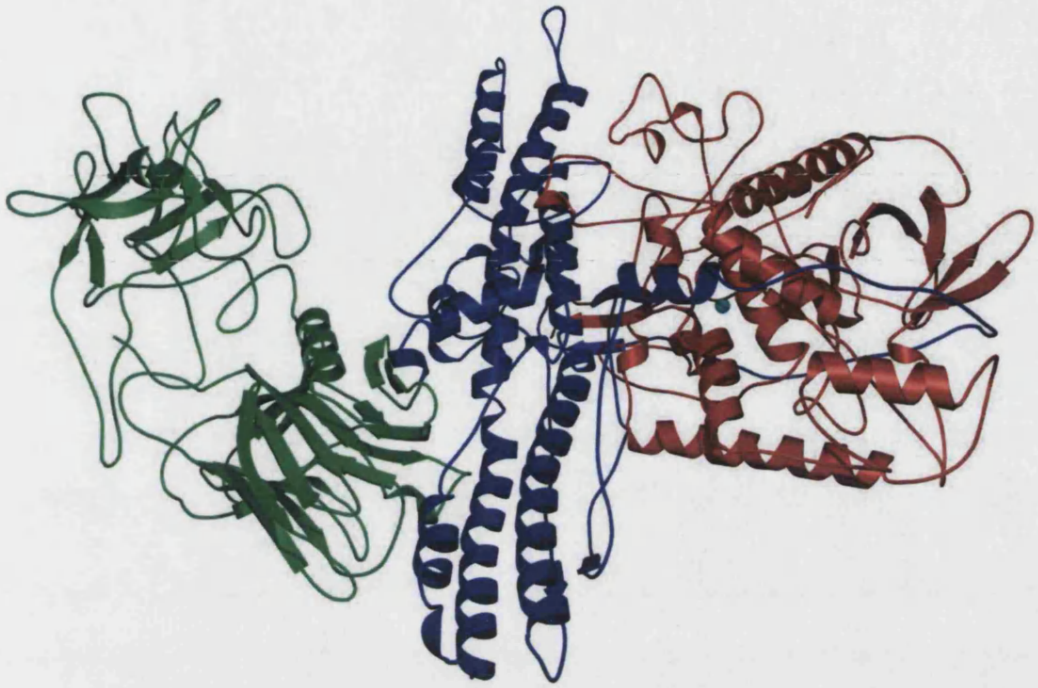
### 4.1.3 CNT Structure

The crystal structures of BoNT/A (Lacy et al. 1998) (Figure 33), BoNT/B (Hanson and Stevens 2000; Swaminathan and Eswaramoorthy 2000b) and TeNT-$H_C$ (Umland et al. 1997) have been determined. The BoNT/A and BoNT/B holotoxins are structurally similar at the tertiary level and there are similarities between the structures of the $H_C$ domains of the three toxins. The three functional domains (LC, $H_N$ and $H_C$) in BoNTs A and B are structurally distinct and arranged in a linear fashion, such that there are no contacts between the catalytic and binding domains.

#### 4.1.3.1 Catalytic Domain (LC)

The catalytic domains of the seven BoNTs share up to 36% sequence identity and the light chains of TeNT and BoNT/B are over 50% identical. The LC comprises a mixture of α-helix and β-strand elements and acts as a zinc-dependent endopeptidase. The active site, located in a deep cleft on the LC surface, is accessible via a channel and contains a zinc ion (or two, in the case of BoNT/C1) and a conserved zinc-binding motif (HExxH) that is characteristic of all zinc-dependent endopeptidases.

**Figure 33 - Crystal Structure of Botulinum Neurotoxin A**



The crystal structure of BoNT/A (PDB code 3BTA; Lacy et al. 1998) revealed that the three functional domains are structurally distinct and arranged in a linear fashion such that there are no contacts between the catalytic and binding domains. The catalytic domain (depicted in red) contains the catalytic zinc ion (in cyan). The translocation domain (in blue) has two unique structural features: The first is a long loop, known as the translocation belt, that wraps around the catalytic domain and the second is the pair of long α-helices that run anti-parallel to, and wrap around, each other. The binding domain (in green) is divided into two sub-domains, one of which adopts a β-trefoil fold and the other with jelly-roll topology, and is oriented such that the C-terminal sub-domain makes no contacts with the central translocation domain. This figure was prepared using Molscript, Povray and GIMP.

The $Zn^{2+}$ is coordinated by the imidazole rings of the two histidines of the HExxH motif and a water molecule bound to the glutamic acid (Pellizzari et al. 1999). In addition, the crystal structure of BoNT/A (Lacy et al. 1998) revealed the presence of a second glutamic acid and a tyrosine close to the zinc. These residues are conserved among the CNTs and mutation of this tyrosine in TeNT to alanine abolishes catalytic activity (Rigoni et al. 2001). Based on structural similarity between their active sites, the catalytic mechanism of BoNT/A is believed to be similar to that of thermolysin (in which $Zn^{2+}$ is coordinated by four ligands, the histidines of the HExxH motif, a glutamic acid and one water molecule). Substrate binding by thermolysin displaces the water

molecule and proteolysis proceeds via a general-base-type mechanism in which the unprotonated imidazole ring of one histidine functions as a base, hydrating the metal-activated substrate (Mock and Stanford 1996).

Site-directed mutagenesis studies identified active site residues involved in the catalytic mechanism of BoNT/A. Substitution of $Glu^{224}$ (the glutamic acid of the HExxH motif) with glutamine abolished catalytic activity, whilst substitution with aspartic acid retained ~1% of the activity of the native toxin (Li et al. 2000). This indicates that the carboxyl group of $Glu^{224}$ is essential for proteolysis (in the proposed mechanism, it coordinates the structural water molecule) and that its location is precisely positioned in the active site. $Glu^{271}$ and $Tyr^{375}$ in TeNT (equivalent to $Glu^{261}$ and $Tyr^{365}$ in BoNT/A) are also essential for proteolysis (Rossetto et al. 2001) and are involved in transition state stabilisation (Binz et al. 2002). Evidence for this was first provided by analysis of the crystal structure of BoNT/B in complex with synaptobrevin II (Hanson and Stevens 2000) where loop rearrangements upon substrate binding were observed to bring $Arg^{369}$ and $Tyr^{372}$ into closer proximity with the scissile bond of the substrate.

#### 4.1.3.2 Translocation Domain ($H_N$)

The central translocation domain has two unique structural features, the most prominent of which is a pair of long, amphipathic α-helices that run anti-parallel to, and wrap around, each other. The other is a long loop, known as the translocation belt, that wraps around the catalytic domain. In BoNT/A, the active site cleft is shielded by the translocation belt but in BoNT/B, this loop is shorter and occupies a different position, exposing the active site (Hanson and Stevens 2000).

The heavy chains of BoNTs A and B form ion-conducting channels that span artificial membranes (Blaustein et al. 1987; Donovan and Middlebrook 1986; Schmid et al. 1993). The channels gate cations and are maximally active at acidic pH (Sheridan 1998). Channel-forming activity has been assigned to the CNT-$H_N$, although reduction of the interchain disulphide bond removes the ability of the BoNT/B to form channels in phospholipid bilayers (Flicker et al. 1999), indicating that structural elements from the other domains may also be required. The pore-forming domains of other bacterial toxins share a common structural motif that is not observed in BoNT/A (Lacy et al. 1998). Instead, the tertiary structure of BoNT/A-$H_N$ more closely resembles coiled-coil viral proteins, which have an acid-induced ability to undergo conformational changes and penetrate membranes but do not translocate through pores. However, the long helices in BoNT/A-$H_N$ lack the heptad repeat that is associated with true coiled coils.

The candidate transmembrane region in BoNT/A (residues 659-681) increases the permeability of lipid bilayers (Oblatt-Montal et al. 1995). The transmembrane sections of pore-forming proteins tend to be α-helical but in the crystal structures of BoNT/A (Lacy et al. 1998) and BoNT/B (Swaminathan and Eswaramoorthy 2000b), the putative membrane-spanning regions do not adopt a helical conformation. Furthermore, channel formation by CNTs is pH-dependent and the residues in BoNT/A most likely to titrate at low pH, $His^{551}$ and $His^{560}$, are located in a loop connecting the translocation belt to the main body of the $H_N$ domain (Lacy et al. 1998). Taken together, these observations suggest that some conformational changes must take place in $H_N$ leading to channel formation.

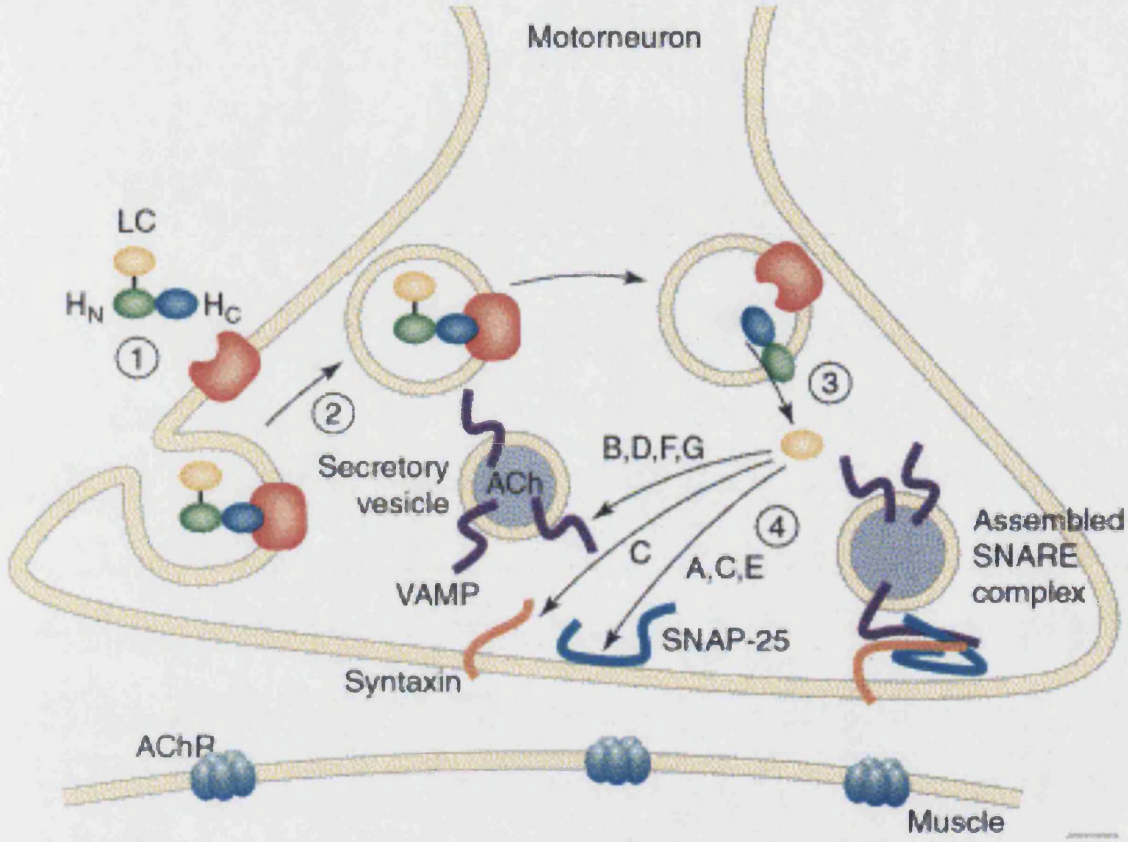
#### 4.1.3.3 Binding Domain ($H_C$)

CNTs are structurally homologous with respect to their binding domains but sequence identity among them is low. The $H_C$ has two sub-domains of roughly equal size, both composed of mainly β-sheet secondary structure, oriented such that there are no contacts between the C-terminal sub-domain and $H_N$ (Lacy et al. 1998). The $H_C$ facilitates the first step in the CNT mechanism of action by recognising and binding to ganglioside and protein receptors on presynaptic nerve termini. The jelly-roll topology of the N-terminal sub-domain is similar to the structure of many lectins but mutations in the C-terminal half of TeNT-$H_C$, which adopts a β-trefoil fold, have been shown to affect ganglioside binding. A stretch of 34 amino acids at the carboxyl terminus of TeNT (residues 1282-1315) have been identified as sufficient for ganglioside binding (Shapiro et al. 1997) and co-crystallisation of TeNT with carbohydrates identified areas on the toxin surface that bind sugars (Emsley et al. 2000).

Sequence-to-structure alignment of the CNTs (Ginalski et al. 2000) revealed a set of five conserved residues within the C-terminal sub-domain of $H_C$. Located in a shallow cleft on the TeNT-$H_C$ surface, $Asp^{1222}$, $Ser^{1287}$, $Trp^{1289}$, $Tyr^{1290}$ and $Gly^{1300}$ are postulated to form the ganglioside-binding motif. Site-directed mutagenesis has since proved that $Tyr^{1290}$ is required for ganglioside binding and TeNT activity (Sutton et al. 2001). Furthermore, the motif $S^{1287}XWY^{1290}...G^{1300}$ has been shown to form a sugar-binding pocket in conjunction with residues $Asn^{1219}$, $Asp^{1222}$ and $His^{1271}$ (Rummel et al. 2003). The crystal structure of BoNT/B in complex with sialyllactose revealed that the sugar binds in a cleft between $Trp^{1261}$ and $His^{1240}$, forming hydrogen bonds with $Glu^{1188}$, $Glu^{1189}$, $His^{1240}$ and $Tyr^{1262}$ and hydrophobic interactions with $Trp^{1261}$ (Swaminathan and Eswaramoorthy 2000b). However, the sugar did not interact with residues $Lys^{1267}$, $Arg^{1268}$, $Lys^{1269}$ or $Glu^{1265}$, as was proposed for TeNT (Shapiro et al. 1997).

### 4.1.4 Mechanism of Action

**Figure 34 - Mechanism of Action of Botulinum Neurotoxins**




Cartoon outlining the steps involved in the mechanism of action of Botulinum Neurotoxins (this diagram is reproduced from Turton et al (2002) Trends Biochem Sci, **27**, 552-558; see Appendix 3).

The toxin recognises and binds to protein and ganglioside receptors on the surface of the neuron via the $H_C$ domain (in blue). Once a ternary complex has been formed, the toxin is taken up by receptor-mediated endocytosis into an endosome-like compartment. The low pH of the endosome interior induces structural rearrangement of the $H_N$ domain (in green), culminating in the formation of an ion-conducting channel that spans the endosomal membrane. The catalytic light chain (in yellow) translocates through the pore of this channel and is released from the toxin heavy chain by reduction of the interchain disulphide bond. Inside the cytosol, the light chain interacts with and cleaves its target SNARE protein (syntaxin, synaptobrevin or SNAP-25), thereby inhibiting the release of neurotransmitter into the synapse.

The mechanism of action of BoNT can be divided into four stages (Figure 34). The toxin binds to specific receptors on the neuron surface and is internalised into an endosomal compartment. The low pH of the endosome interior induces structural rearrangement enabling the toxin to form an ion channel spanning the endosomal membrane. The catalytic domain translocates through the ion channel and, following reduction of the interchain disulphide bond, moves through the neuron cytoplasm to act

upon its target protein. The targets for all CNTs are the soluble NSF-attachment protein receptors (SNAREs) (see section 4.1.4.3). The roles of receptor-mediated endocytosis, pH-dependent translocation and the zinc ion in the mechanism of action of BoNT/A have been experimentally confirmed (Kalandakanond and Coffield 2001).

### 4.1.4.1 Recognition and Internalisation of BoNT by Neurons

BoNTs bind to gangliosides on presynaptic nerve termini via their $H_C$ domains. Gangliosides are glycosphingolipids that are abundant in neuronal cell membranes. They contain a common core to which at least one sialic acid residue is bound. GM1 and GD1b gangliosides have one and two sialic acids, respectively, attached to the internal galactose of the core, whilst GT1b and GQ1b gangliosides have one and two sialic acids bound to the terminal galactose. GD1b and GT1b gangliosides, which have the highest affinity for CNTs, are thought to be present only at synapses, accounting for the high specificity of BoNTs for nerve cells (Ahnert-Hilger and Bigalke 1995).

However, the affinity of BoNTs for GD1b and GT1b gangliosides is too weak to account for their potent toxicity. The binding affinity also varies with toxin serotype, type of ganglioside, temperature, pH and ionic conditions (Ahnert-Hilger and Bigalke 1995). In addition, some birds with a high content of polysialogangliosides in their neuronal cell membranes are resistant to TeNT (Montecucco 1986). Taken together, these data suggest that the mechanism of CNT recognition by neurons is not solely dependent upon gangliosides. A "double-receptor model" has been proposed (Montecucco 1986) whereby CNTs must also bind to a specific protein receptor on the neuron surface. The activity of BoNT/A has since been shown to require the presence of gangliosides and synaptotagmin I in neuronal cell membranes (Yowler et al. 2002) and BoNT/B recognises a complex of GT1b/GD1a gangliosides and synaptotagmin II on reconstituted lipid vesicles (Kozaki et al. 1998).

Peripheral cholinergic neurons are the major target for BoNTs, although CNS neurons are also susceptible to intoxication (Verderio et al. 1999). Once a ternary complex comprising the toxin and its receptors has been formed, BoNT is internalised by the neuron into an acidic membrane-bound compartment. Internalisation is thought to occur via receptor-mediated endocytosis involving clathrin-coated vesicles because the uptake of BoNT/A can be inhibited by the application of drugs that inhibit endocytosis (Simpson 1988). After internalisation, BoNT can no longer be neutralised by anti-toxin antibodies (Montecucco et al. 1994).

#### 4.1.4.2 Channel Formation and LC Translocation

In common with other bacterial toxins, BoNT is thought to undergo structural rearrangement inside an acidic compartment within the neuron. Diphtheria toxin (DT) enters cells via receptor-mediated endocytosis and its catalytic domain is transported to the cytosol across the endosomal membrane following a channel-formation step induced by the low pH of the endosome interior (Donovan et al. 1981; Kagan et al. 1981). Following internalisation by the neuron, BoNT is also contained within an endosome and the low pH environment is thought to induce structural changes leading to channel formation (Montecucco and Schiavo 1995; Pellizzari et al. 1999). Treatment of neuronal cells with methylamine hydrochloride (a compound that blocks endosome acidification) antagonises the action of BoNT/A, indicating that the toxin must be located within an acidic compartment within the neuron at some point during its mechanism of action (Kalandakanond and Coffield 2001). Electron microscopy has confirmed the formation of pores by BoNT/A in vesicle membranes (Schmid et al. 1993).

The pores of channels formed by DT (~18Å in diameter) are large enough for the toxin's catalytic domain to pass through (Kagan et al. 1981). Similarities noted among the channel-forming properties of BoNTs, TeNT and DT (Blaustein et al. 1987) have led to the proposal that the CNT LC passes through the channel formed by the toxin heavy chain (Hoch et al. 1985). Channel formation by all these toxins requires low pH on the "endosomal" side of the membrane, compared with neutral pH on the other, and both channel activity and conductance increase when the pH on the "cytosolic" side of the membrane is elevated. In addition, channel formation by BoNT/B requires the interchain disulphide bond to be intact (Flicker et al. 1999), indicating that the light and heavy chains of BoNT do not dissociate until transcytosis has taken place. However, there is some evidence to suggest that the LC occludes the heavy-chain channel and that reduction of the interchain disulphide bond unblocks the $H_N$ channel (Koriazova and Montal 2003).

Circular dichroism spectroscopy revealed that the BoNT/A-LC could only move through the channel formed by the heavy chain in an unfolded conformation (Koriazova and Montal 2003). This is in agreement with the finding that the diameter of channels formed by BoNTs in artificial membranes is only ~8Å (i.e. too small to accommodate a fully-folded 50kDa globular protein) (Hoch et al. 1985). The structural changes induced in BoNT/A-LC by low pH are completely reversible and do not affect zinc binding (Li and Singh 2000b). Therefore, the mechanism of translocation is postulated to involve partial unfolding of the LC within the endosome followed by

translocation into the cytosol where the LC dissociates from the toxin heavy chain and refolds to an active conformation (Koriazova and Montal 2003).

#### 4.1.4.3 Inhibition of Neurotransmitter Release

Structural changes in BoNT/A-LC upon reduction of the interchain disulphide bond increase enzymatic activity (Cai et al. 1999), suggesting that the structure of the LC is different when it is released from the endosome to its structure in the holotoxin. It is thought that these structural changes increase the accessibility of the catalytic site by orienting key active site residues for optimal activity (Hanson and Stevens 2000). The crystal structure of BoNT/B in complex with synaptobrevin-2 (Hanson and Stevens 2000) revealed rearrangement of three active site loops upon separation of the two chains that bring $Arg^{369}$ and $Tyr^{372}$ into close proximity with the scissile bond of synaptobrevin. Separation of the LC from the toxin heavy chain also relieves the hindrance imposed by the translocation belt, which in the crystal structure of BoNT/A (Lacy et al. 1998) can be seen to occlude the active site cleft.

**Table 9 - Substrate Specificity of the Botulinum Neurotoxins**

| BoNT Serotype | Substrate SNARE | Peptide Bond Cleaved |
|---|---|---|
| A | SNAP-25 | Gln197-Arg198 |
| B | Synaptobrevin II | Gln76-Phe77 |
| C1 | Syntaxin<br>SNAP-25 | Lys253-Ala254<br>Arg198-Ala199 |
| D | Synaptobrevin II | Lys49-Ile50 |
| E | SNAP-25 | Arg180-Ile181 |
| F | Synaptobrevin II | Gln48-Lys49 |
| G | Synaptobrevin II | Ala81-Ala82 |

The targets of BoNTs within the neuron cytosol are the SNARE proteins SNAP-25, VAMP-2 and syntaxin. The toxins show remarkable substrate specificity, despite a high level of homology among their amino acid sequences, in that no two BoNTs cleave the same substrate at the same peptide bond. In addition, the BoNTs show a requirement for substrates of a certain size. This is because the BoNT-LC contains two SNARE recognition sites, one for the nine-residue SNARE motif and one for the scissile peptide bond, and the substrate must bind to both sites for cleavage to occur.

The targets of all CNTs are the SNAREs, proteins that mediate membrane fusion at the nerve terminal and, thus, play an essential role in the release of neurotransmitters into the synapse. BoNTs A and E cleave SNAP-25 (synaptosomal associated protein of 25kDa); BoNT serotypes B, D, F, and G act upon vesicle-associated membrane protein (VAMP-2, also known as synaptobrevin II); and

BoNT/C1 targets both syntaxin and SNAP-25. BoNTs exhibit remarkable substrate specificity in that no two serotypes cleave the same protein at the same peptide bond (Table 9). Not only are the BoNTs highly specific for their substrates in terms of sequence, but they also show a requirement for substrates of a minimum length. BoNT/B, for example, will only cleave substrates comprising at least 30 amino acid residues whilst TeNT and BoNT/A require substrates of at least 50 amino acids (Shone et al. 1993). This is because the BoNT-LC contains two substrate recognition sites and the target SNARE must bind to both for proteolysis to occur (Pellizzari et al. 1996).

The SNARE proteins share a common nine-residue structural motif, known as the SNARE motif, which is thought to adopt a helical conformation (Rossetto et al. 1994). There are at least two copies of this motif in each SNARE – VAMP-2 and syntaxin each contain two copies and SNAP-25 has four copies. It is thought that BoNTs bind the SNARE motif and the scissile peptide bond, which is located further towards the C-terminus in the amino acid sequence than the SNARE motif (Rossetto et al. 1994). This explains the requirement of BoNTs for substrates of a certain length. Furthermore, residues located N-terminal to the cleavage site in VAMP-2 are known to be important for substrate recognition by BoNT/B (Wictome et al. 1996).

#### 4.1.4.4 Role of SNARE Complex Assembly in Neurotransmission

The release of neurotransmitters into synapses requires the formation of a trimeric complex by the SNARE proteins syntaxin, synaptobrevin and SNAP-25. Both syntaxin and SNAP-25 reside in the plasma membrane at the nerve terminal, whilst synaptobrevin is present in synaptic vesicle membranes. Syntaxin and SNAP-25 are localised in overlapping regions of the plasma membrane, generating a high local concentration of SNAREs for efficient membrane fusion (Lang et al. 2001). These "clusters" are distinct from the protein-containing lipid rafts that are involved in toxin internalisation. Formation of the trimeric "SNARE complex" brings synaptic vesicles close to the plasma membrane at the nerve terminal, facilitating membrane fusion and the release of neurotransmitters in response to $Ca^{2+}$ influx. Disruption of the $Ca^{2+}$-dependent triggering of exocytosis by CNTs indicates that SNARE complex assembly occurs immediately prior to membrane fusion (Gerona et al. 2000).

The SNARE complex is stabilised by the coiled-coil interactions of four α-helices from the three SNAREs (Poirier et al. 1998). Although sequence similarity is low between the SNAREs, a pattern of conserved hydrophobic amino acids arranged in heptad repeats has been observed (Weimbs et al. 1997). The first and fourth residues of each heptad are hydrophobic, forming an α-helix with a hydrophobic face (Weis and

Scheller 1998). The SNARE complex assembly brings together the hydrophobic faces of four helices, one each from syntaxin and synaptobrevin and two from SNAP-25 (Jahn and Sudhof 1999). Pairs of SNAREs bind each other weakly, but their association is greatly enhanced when all three proteins are present (Hayashi et al. 1994). The stability of the four-helix bundle is extremely high (Fasshauer et al. 1998) and the SNARE complex must be actively disassembled by the ATPase N-ethylmaleimide-sensitive factor (NSF) and the soluble NSF-attachment protein α-SNAP following neurotransmitter release. The cleavage of SNAREs by CNTs does not prevent SNARE complex assembly, but results in the formation of a non-functional complex that cannot couple $Ca^{2+}$ influx with neurotransmission (Humeau et al. 2000).

Synaptobrevin is protected from proteolysis by CNTs when it forms a complex with the other two SNAREs or with NSF and α-SNAP (Pellegrini et al. 1994). VAMP-2 was found to bind to BoNT/B in a random coil conformation (Hanson and Stevens 2000), but must be helical for the formation of a functional SNARE complex. SNARE complex assembly must therefore protect the scissile bond and/or SNARE motif in each SNARE from recognition by CNTs. It has also been suggested that the high entropic cost associated with helix formation might lead to the preferential binding of uncomplexed VAMP to BoNT/B, which binds it in a random coil conformation (Hanson and Stevens 2000). If this is true, SNAREs will bind to CNTs in preference to forming functional SNARE complexes in intoxicated neurons.

#### 4.1.4.5 Mechanism of Action of TeNT

Unlike BoNTs, TeNT enters the neuron via synaptic vesicle recycling (Matteoli et al. 1996). Inhibition of ganglioside synthesis has been shown to protect neurons against TeNT activity (Williamson et al. 1999). The high sensitivity of neurons to TeNT requires lipid rafts and at least one protein receptor (Munro et al. 2001), but the identity of this 15kDa glycoprotein receptor remains elusive (Herreros et al. 2000). After internalisation, TeNT moves through the neuron by retrograde axonal transport and is transcytosed to the spinal cord inhibitory interneurons where it cleaves its target SNARE VAMP-2 (Verderio et al. 1999). The resultant spastic paralysis is due to the loss of spinal inhibitory control over motorneuron activity.

TeNT cleaves VAMP-2 at the same peptide bond as BoNT/B (Pellegrini et al. 1994). It has been shown that VAMP-2 is not the major substrate for TeNT in adrenal chromaffin cells because the inhibition of neurotransmission by TeNT in these cells is not attenuated by the application of captopril (a metallopeptidase inhibitor) (Hohne-Zell et al. 1993). TeNT can activate a calcium-dependent transglutaminase (Facchiano and

Luini 1992) and it has since been postulated that TeNT inhibits neurotransmission by the proteolysis of both synaptobrevin and the transglutaminase (Ashton et al. 1995).

### 4.1.5 Medical Applications for Botulinum Neurotoxins

Preparations of BoNT/A have been licensed for use as pharmaceuticals since 1989, when it was shown to be an effective treatment for strabismus, hemifacial spasm and blepharospasm (Johnson 1999). The preparation Botox is perhaps most famous for its use as a non-surgical cosmetic treatment for removing facial wrinkles. The high specificity of BoNT/A for motorneurons combined with its longevity of action (up to 4-6 months) make it a useful biopharmaceutical and its use has now been extended to a variety of neuromuscular disorders and an increasing number of other syndromes.

#### 4.1.5.1 Therapeutic Utility of BoNT/A

Local injections of the toxin weaken overactive muscles, providing relief from spasticity. BoNT/A is now well established as a treatment for several ophthalmologic disorders, such as strabismus, and movement disorders, including writer's cramp and torticollis (Munchau and Bhatia 2000). BoNT/A has also been tested as a potential therapeutic agent for many more muscular disorders, such as tic disorders, tremor and spasticity after stroke. In addition to its effects at the neuromuscular junction, BoNT/A also acts upon cholinergic autonomic neurons (Hambleton 1992) and may, therefore, be used for the treatment of a variety of autonomic disorders, such as anal fissure (Naumann et al. 1999). BoNT/A is a useful therapy for hyperhidrosis, which is caused by sympathetic neurotransmission in sweat glands (Heckmann et al. 2001). However, many of the potential applications of BoNT/A to the autonomic nervous system remain to be thoroughly tested.

BoNT/A is currently being investigated as a treatment for chronic pain. In fact, BoNT/A in combination with physiotherapy has been demonstrated to provide better relief from myofascial pain syndrome than conventional steroid treatment (Porta 2000). BoNT/A can also be an effective treatment for migraine with the added advantage of reducing migraine severity, frequency and associated side effects (Silberstein et al. 2000).

#### 4.1.5.2 Side-effects Associated with BoNT/A Therapy

As with all pharmaceuticals, there are several side effects and complications associated with the therapeutic use of BoNT/A. There is a need for repeated injections because the paralysis induced only lasts for 4-6 months, but the main problem with the therapy is the development of resistance by the production of anti-BoNT/A antibodies.

Reducing the dose and frequency of injections can limit antibody production, although resistance to BoNT/A can be overcome by the use of alternative BoNT serotypes. BoNT/C1 has been demonstrated to have a similar effect and longevity of action to BoNT/A (Eleopra et al. 1997) and may become a valid alternative therapy. Another side effect is that the toxin can diffuse from the site of injection into neighbouring muscles, causing transient ptosis (Johnson 1999). This can be debilitating (e.g. dysphagia) and as long lasting as the duration of action of the toxin. Whilst no allergic reactions to BoNT/A have been reported, other side effects include symptoms of influenza and brachial plexopathy (Munchau and Bhatia 2000). The therapeutic use of BoNT/A in children is carefully monitored and it is contra-indicated during pregnancy and while breast-feeding.

#### 4.1.5.3 Retargeting the Endopeptidase Activity of BoNT/A

Non-toxic fragments of TeNT-$H_C$ have been proposed as delivery vehicles for specifically targeting ligands to peripheral or CNS neurons because the CNT binding domain is highly specific for cholinergic nerve terminals. The heavy chains of TeNT, BoNT/A and BoNT/B have recently been used to target model compounds to human neuroblastoma cells (Goodnough et al. 2002; Zdanovskaia et al. 2000). However, the potent endopeptidase activity of these toxins can potentially be retargeted to a variety of neuronal tissues if the $H_C$ domain is replaced by a non-native cell-binding moiety with an alternative specificity.

A fragment of BoNT/A comprising only the LC and $H_N$ domains ($LH_N$/A) has been recombinantly expressed and purified (Chaddock et al. 2002). $LH_N$/A retains the full catalytic activity of the BoNT/A holotoxin but is considered non-toxic because it lacks a cell-binding domain. $LH_N$/A has been chemically coupled to nerve growth factor (Chaddock et al. 2000a), wheat germ agglutinin (Chaddock et al. 2000b) and *Erythrina cristagalli* lectin (Duggan et al. 2002) and specifically retargeted *in vitro* to cells of neuronal and non-neuronal origin. The conjugate comprising $LH_N$/A and ECL (known as $LH_N$/A-ECL) exhibits an *in vitro* selectivity for nociceptive afferents and is of potential therapeutic use in the treatment of chronic pain (Duggan et al. 2002). As described in Chapter One, $LH_N$/A-ECL has equivalent potency and longevity of action within DRG to the BoNT/A holotoxin in its target neurons and a catalytically inactive mutant of this fragment ($LH_N$/A($H^{227}Y$)) has been used as a negative activity control.

## 4.2 MATERIALS AND METHODS

### 4.2.1 Materials

$LH_N/A$ and $LH_N/A(H^{227}Y)$ were provided by Dr J.A. Chaddock (Health Protection Agency, UK). The toxin fragments (873 amino acids in length, including a two amino acid N-terminal extension arising from expression) were expressed and purified by the Toxin Therapeutics Division at CAMR as previously reported (Chaddock et al. 2002) and briefly described in Appendix 4. The purified protein stocks were stored at -20°C in a buffer comprising 100mM sodium chloride and 20mM HEPES pH 7.0. For crystallisation trials, the protein stocks were diluted to 7-13mg/ml with milliQ water.

The electrophoresis system was purchased from BioRad (UK) and High and Low Molecular Weight Markers were obtained from Pharmacia (UK). Slide-A-Lyzer dialysis kits were purchased from Pierce (USA). Glycerol and Tris were obtained from VWR International. Sodium dodecyl sulphate, HEPES, bromophenol blue, citric acid and glycine were purchased from Sigma. Methanol, sodium chloride and acetic acid were purchased from Fisher Chemicals. All buffers were prepared with milliQ water.

Reagents used to prepare crystallisation solutions were of high quality and purity. All crystallisation solutions were made with AnalaR water (VWR International). Crystallisation plates and Structure Screen I and II were purchased from Molecular Dimensions Ltd. Glass coverslips, Crystal Screen and a PEG/Ion matrix were purchased from Hampton Research and JBScreen 7 from JenaBioScience. PEG 4000, PEG 5000 MME, PEG 20000, tert-butanol, dioxane, potassium thiocyanate, sodium sulphate and malate were purchased from Fluka. PEG 6000, isopropanol, ammonium chloride and glycerol were obtained from VWR International. Potassium phosphate, sodium phosphate, acetone and ethanol were purchased from Fisher Chemicals. Potassium chloride was obtained from Fissons (UK). All other reagents were purchased from Sigma.

### 4.2.2 SDS-PAGE

Samples from the $LH_N/A$ and $LH_N/A(H^{227}Y)$ protein stocks were diluted with milliQ water to a concentration of ~0.2mg/ml and mixed in a 1:1 ratio with sample buffer (20% glycerol, 100mM Tris, pH 6.8, 4% SDS and 0.01% bromophenol blue). High and Low Molecular Weight protein markers were heated for 5 minutes in a heating block. Markers and samples were loaded onto a 5% stacking gel and separated on a 10% gel in running buffer (25mM Tris, 250mM glycine, pH 8.3, and 0.1% SDS) at 12mA until the

dye front ran off the bottom of the gel. The gel was stained overnight in Coomassie Brilliant Blue, then destained with methanol/acetic acid before being photographed.

### 4.2.3 Crystallisation Trials

The hanging drop vapour diffusion method was employed. Crystallisation plates were incubated at 4°C, 16°C or 22°C in temperature-controlled incubators or at room temperature. Drops comprising 1.5µl protein and 1.5µl mother liquor were equilibrated against wells containing 700µl of mother liquor.

#### 4.2.3.1 $LH_N/A$

Randomised sparse matrix crystallisation screens (Crystal Screen and Structure Screens I and II) were set up to sample a wide range of potential crystallisation conditions. The conditions included in the commercial crystallisation screens are listed in Appendix 2. A PEG/Ion matrix was used in conjunction with grid screens, in which the concentrations of PEG 4000, PEG 6000 or ammonium sulphate were varied with pH, because PEG and ammonium sulphate are the precipitants most commonly reported for protein crystallisation. In addition, conditions reported for the crystallisation of BoNT/A were tested (Kadkhodayan et al. 2000; Lacy et al. 1998; Stevens et al. 1991).

A literature search was performed to uncover conditions previously reported for successful crystallisation of toxins. These conditions mainly involved PEG and MPD, so JBScreen 7 (Appendix 2) was also tested. An alcohol/pH screen tested the effect of MPD, tert-butanol, 1,6-hexanediol and isopropanol as precipitants with four buffers in the pH range 4.6-7.0. In addition, isopropanol and alcohol screens based on JBScreens 8 and 9 were set up.

Conditions reported for the crystallisation of CNTs and CNT fragments, that had not previously been used, and a variety of conditions reported for the crystallisation of zinc-dependent proteases (Appendix 4) were set up to cover a range of conditions that may yield crystals of $LH_N/A$. Those conditions producing interesting results were repeated and optimised. The technique of streak-seeding was reported as essential for the growth of diffraction-quality crystals of BoNT/B (Swaminathan and Eswaramoorthy 2000a; Swaminathan and Eswaramoorthy 2000b), so it was applied to $LH_N/A$.

#### 4.2.3.2 $LH_N/A(H^{227}Y)$

Initial crystallisation screens involved Structure Screens I and II, a PEG/Ion matrix and a PEG 6000 grid screen as well as conditions reported for the crystallisation of

BoNT/A. Based upon their reported success as crystallisation agents (McPherson 2001) and the observation that a salt-based condition had produced shiny crystalloids, sodium malonate, sodium tartrate and ammonium formate were tested as precipitation agents. An alcohol/pH comparison, a divalent cation screen and the conditions reported for the crystallisation of CNTs and zinc-dependent proteases were set up, as for $LH_N/A$. Two alcohol-based screens were set up in which the concentrations of MPD and 1,6-hexanediol were varied with increasing pH. Crystallisation trials based on JBScreens 8 and 9 were also set up.

### 4.2.4 Circular Dichroism Spectroscopy

#### 4.2.4.1 Dialysis and Protein Estimation

Samples of the $LH_N/A$ and $LH_N/A(H^{227}Y)$ protein stocks were diluted 1/50 with milliQ water for CD spectroscopy. . A diluted sample of the buffer (100mM sodium chloride and 20mM HEPES, pH 7.0) was also prepared, so that a blank (i.e. protein-free) CD spectrum could be recorded.

200µl samples of $LH_N/A(H^{227}Y)$ were injected into Slide-A-Lyzer cassettes and dialysed overnight at 4°C against 100mM sodium chloride and 20mM sodium HEPES (pH 7.0 buffer) or 100mM sodium chloride and 20mM sodium citrate (pH 5.0 buffer). After dialysis, the protein solutions were diluted 1/50 with milliQ water and the protein concentration of each was determined. A sample of each buffer was also diluted 1/50 and analysed for protein content. The diluted protein and buffer samples were then used in CD experiments.

The concentration of protein in each of the four diluted protein samples was determined using the spectrophotometric method and related to the theoretical molar extinction coefficient using the Beer Lambert law:

$$A_{280} = \varepsilon c l$$

where: $A_{280}$ is the absorbance of UV radiation at 280nm

$\varepsilon$ is the extinction coefficient of the sample at 280nm

c is the concentration of the protein sample

l is the path length

The amount of light (of wavelength 280nm) absorbed by each sample was measured with an Ultrospec 2000 spectrophotometer (Pharmacia Biotech, UK) in a 1cm path-length quartz cuvette.

#### 4.2.4.2 Circular Dichroism Spectroscopy

CD spectra were recorded with a JASCO J-600 spectropolarimeter across the wavelength range 250-185nm. Six scans were recorded, using a quartz cell with a path length of 0.02cm, and averaged for each sample to increase the signal-to-noise ratio. The scan speed was 10nm/min and the response time was fixed at 4s. CD spectra for the buffer blanks were recorded separately (using the same parameters as for the corresponding protein sample) from samples of dialysis buffer diluted 1/50 (so that the final concentration of sodium chloride was the same as in the protein samples).

CD spectra recorded for the protein samples were normalised, by subtracting the appropriate buffer blank, and corrected for protein concentration. CD was then converted to mean residue ellipticity and the spectra were analysed by three different algorithms.

#### 4.2.4.3 Analysis of CD Spectra

CD spectra were deconvoluted and secondary structures assigned using the CONTINLL, SELCON3 and CDSSTR programs available through the DICHROWEB server (http://public-1.cryst.bbk.ac.uk/cdweb/html; Lobley and Wallace 2001; Lobley et al. 2002). Two protein reference sets were used in conjunction with each analysis method: Reference set 3 comprises 37 proteins and reference set 6 comprises 42 proteins, five of which are denatured (Sreerama et al. 2000; Sreerama and Woody 2000). Both reference sets cover the wavelength range 185-240nm, which closely matches the wavelength range across which the experimental data were collected.

## 4.3 RESULTS

### 4.3.1 Crystallisation of $LH_n/A$ and $LH_n/A(H^{227}Y)$

Crystallisation trials with $LH_N/A$ and $LH_N/A(H^{227}Y)$ produced similar results. As the two toxin fragments differ in primary structure by only one amino acid, it was expected that they would crystallise under the same, or very similar, conditions. However, despite setting up a large number of crystallisation trials, involving several hundred conditions, no crystals of $LH_N/A$ or $LH_N/A(H^{227}Y)$ were grown. The hanging drops were studied regularly under a polarising light microscope and optimisation of the crystallisation conditions was based upon the appearance of the drops. A number of clear drops were observed, but several of the crystallisation conditions tested produced precipitation or phase separation. After several months, large crystals were observed in three drops, but upon investigation by X-ray diffraction, they were discovered to be composed of salt.

An unusual phase separation effect was observed with five different mother liquors for $LH_N/A$:

- 35% PEG 4000, 0.2M ammonium sulphate and 0.1M sodium acetate, pH 5.6
- 2.5M 1,6-hexanediol, 0.01M manganese chloride and 0.1M sodium citrate, pH 5.6
- 2.0M sodium chloride and 10% PEG 6000
- 35% dioxane
- 1.5M sodium chloride and 10% ethanol

The shiny crystalloids were birefringent under polarised light and it was suggested that they might be some kind of liquid-crystal (therefore, they were described as "shiny crystalloids"). Conditions leading to the formation of shiny crystalloids were repeated and expanded in an attempt to grow crystals suitable for diffraction. It was observed that the removal of manganese chloride from mother liquor comprising 1,6-hexanediol, manganese chloride and HEPES, pH 7.0 produced smaller crystalloids, so a screen was set up to determine the effect of five divalent cations on crystalloid quality. Increasing concentrations of zinc chloride, calcium chloride, manganese chloride, magnesium chloride and cadmium chloride were assayed in conjunction with 2.0M 1,6-hexanediol at pH 7.0. Phase separation and shiny crystalloids were observed in several drops containing MPD or hexanediol in the pH range 5.6-8.5. It was noted that alcohols as precipitants seemed to produce more promising results than mother liquors containing PEG or inorganic salts. Although no crystals had been grown, in addition to shiny crystalloids, small, round, crystal-like structures (referred to as "microcrystals") were observed with several crystallisation conditions. As these were thought to be promising results, all conditions giving rise to microcrystals were repeated and expanded. Mother liquors used for seeding were those that had produced microcrystals and/or crystalloids. Further optimisation of conditions based upon alcohols and PEG were tested, as these precipitants produced promising results.

Initial crystallisation trials involving $LH_N/A(H^{227}Y)$ did not produce any crystals, but shiny crystalloids were grown from four different mother liquors:

- 25% PEG 4000, 0.2M ammonium sulphate and 0.1M sodium acetate, pH 5.6
- 20% isopropanol, 0.2M tri-sodium citrate and 0.1M HEPES, pH 7.5
- 2.0M sodium chloride, 0.1M sodium phosphate monobasic, 0.1M potassium phosphate monobasic and 0.1M MES, pH 6.5
- 2.5M 1,6-hexanediol, 0.01M manganese chloride and 0.1M sodium citrate, pH 5.6

As observed for $LH_N/A$, three of these conditions involved PEG or an alcohol as a precipitant, but one salt-based condition also produced shiny crystalloids. These four conditions were repeated and expanded. No crystals were obtained, but more

shiny crystalloids were grown in several conditions comprising 2.5-3.2M 1,6-hexanediol and 0.01-0.02M manganese chloride in the pH range 5.6-7.5. As with $LH_N/A$, mother liquors containing alcohols appeared to produce more promising results than salt-based conditions. Although no crystals of $LH_N/A(H^{227}Y)$ were yielded from these experiments, MPD and isopropanol seemed to produce the best results, so further optimisation was based on these two precipitants.

#### 4.3.1.1 Analysis of Hanging Drops

Clear drops indicate that the protein and/or precipitant concentration is too low for nucleation and crystallisation (Zeelen 1999). The protein stock was used at a concentration of 7-13mg/ml in a 1:1 ratio with crystallisation buffer in the hanging drops, giving a final protein concentration in each drop of ~5mg/ml. As many different crystallisation conditions were initially tested, those leading to clear drops were not repeated or optimised and the focus of subsequent crystallisation experiments was shifted to those drops in which precipitates or phase separation were observed.

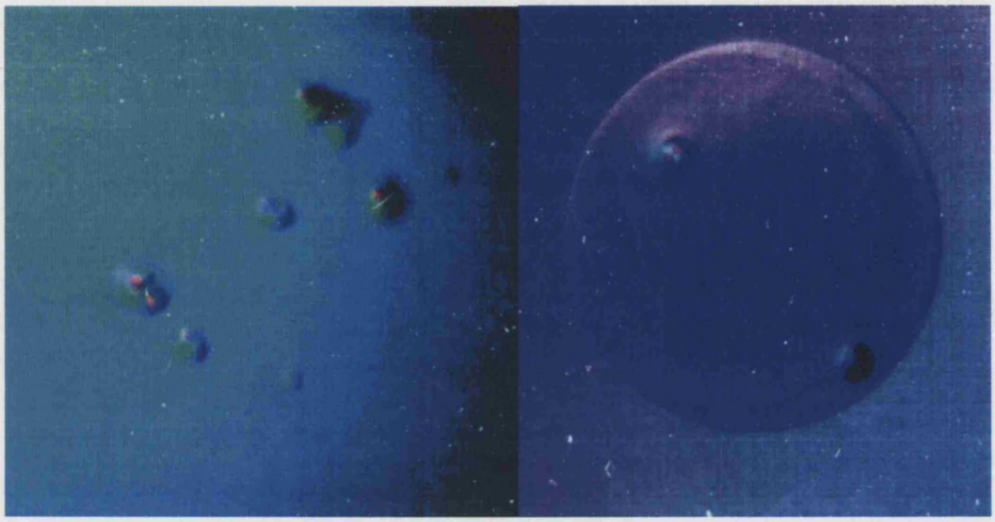
The cause of protein precipitate in a drop might be that the mother liquor does not favour crystallisation (Zeelen 1999). Alternatively, as is the case for clear drops, the concentration of protein and/or precipitant may be too low. The production of a gelatinous precipitate that is dark in colour, however, may be a good starting point for the optimisation of crystallisation conditions. Several of the conditions producing such precipitates were repeated and optimised, by changing the pH, buffer, precipitant and by adding some additive compounds. Different types of phase separation were also observed in many of the hanging drops. Whilst this can be another good starting point for optimisation, it is also possible for crystals to grow in the highly concentrated protein phase of the drop. The conditions leading to phase separation were optimised, by changing the concentration and nature of the precipitant, changing the buffer and pH and by the addition of some additives.

The incubation temperature may also have influenced the success of the crystallisation trials. Many conditions reported for the crystallisation of toxins were incubated at 20-24°C. A number of experiments on $LH_N/A$ and $LH_N/A(H^{227}Y)$ were set up at room temperature, with the crystallisation plates stored inside polystyrene boxes in the laboratory. After several months, when no crystals were observed, it was feared that the temperature inside the boxes might have fluctuated with temperature changes in the laboratory. Changes in temperature alter the degree of supersaturation in the drop and, thus, can hinder nucleation and crystal growth (Drenth 1999). To try to overcome any adverse effects that might be caused by temperature fluctuations, a

refrigeration incubator was set to 22°C and several of the promising conditions were repeated, but no crystals were yielded. Crystallisation plates were also incubated at 4, 8 or 16°C, but these experiments too were unsuccessful.

#### 4.3.1.2 Shiny Crystalloids and Microcrystals

**Figure 35 - Shiny Crystalloids**



This unusual form of phase separation (dubbed "shiny crystalloids") was observed in several of the drops equilibrated against mother liquors containing high concentrations of alcohol as a precipitant. The shiny crystalloids are birefringent when viewed under polarised light and might be some kind of liquid-crystal. However, attempts to optimise the conditions producing these crystalloid structures to yield true crystals of $LH_N/A$ and $LH_N/A(H^{227}Y)$ were unsuccessful.

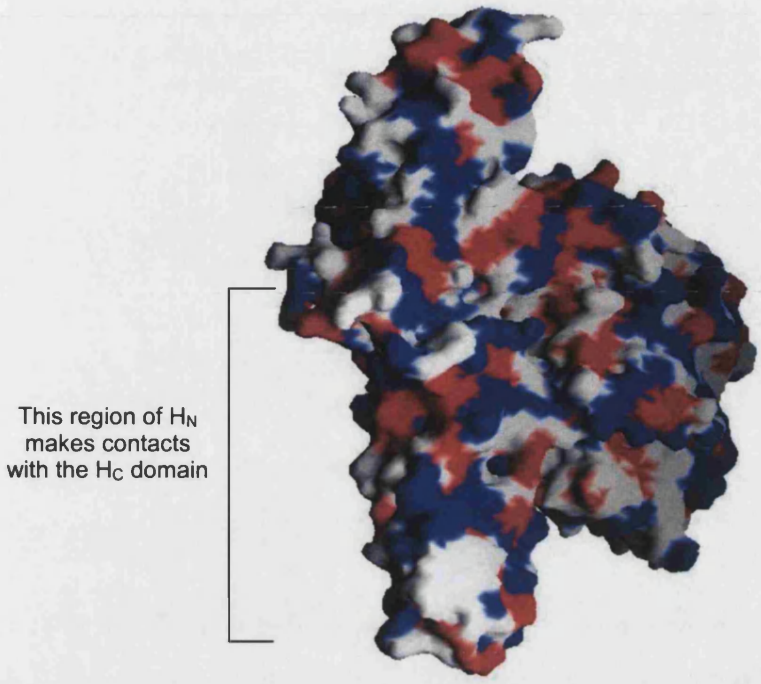
An unusual form of phase separation, dubbed "shiny crystalloids" due to their birefringence under polarised light (Figure 35), was observed for both $LH_N/A$ and $LH_N/A(H^{227}Y)$ under a variety of crystallisation conditions. Shiny crystalloids were not considered to be true crystals, but rather some kind of liquid-crystal. The conditions that generated them were repeated and expanded in order to try to grow true crystals. Most of the crystallisation media producing shiny crystalloids involved alcohols, such as MPD, 1,6-hexanediol and isopropanol, at quite high concentrations as precipitants. It was, therefore, thought that alcohol-based solutions would be more likely to lead to crystals of $LH_N/A$ and $LH_N/A(H^{227}Y)$ than salt or PEG-based conditions. The nature and concentration of the alcohol was varied, as was the pH and nature of the buffer involved, but no crystals were yielded. Substitution of PEG 500 for MPD as a precipitant and the use of PEG 500 as an additive were also ineffective.

In addition, several further crystallisation conditions produced another form of phase separation, termed "microcrystals". These structures were thought to be very small protein crystals, but as they were not observed to grow, it is possible they may represent another form of crystalline precipitate. The conditions producing microcrystals also mostly involved alcohols as precipitation agents. Optimisation was, therefore, based upon the strategy used for optimising the conditions producing shiny crystalloids. However, despite setting up a large number of further experiments, no crystals of $LH_N/A$ or $LH_N/A(H^{227}Y)$ were grown.

Streak seeding was reported to be necessary for growing crystals of BoNT/B (Swaminathan and Eswaramoorthy 2000a). Initial crystallisation screens produced thin, plate-like crystals of BoNT/B, but it was only by seeding with cat whiskers that diffraction-quality crystals could be obtained. In the absence of plate or needle crystals in the $LH_N/A$ and $LH_N/A(H^{277}Y)$ crystallisation trials, shiny crystalloids and microcrystals were used as seeds for macroseeding and streak-seeding experiments. Macroseeding involved the removal of one or two microcrystals from a drop and transfer (via a washing step) to a fresh, pre-equilibrated drop containing precipitant at a slightly lower concentration to that used for microcrystal growth. Streak-seeding, on the other hand, was performed by touching a cat whisker onto some shiny crystalloids (to pick up some crystal seeds) and streaking the seeded whisker across a fresh, pre-equilibrated drop, again with a lower concentration of precipitant than that used to grow the crystalloids. However, neither macroseeding nor streak-seeding yielded crystals of $LH_N/A$.

As $LH_N/A$ and $LH_N/A(H^{227}Y)$ comprise only two of the three functional domains of BoNT/A, it was thought that there might be a large hydrophobic face on one side of $H_N$ that would be protected by the $H_C$ domain in the holotoxin. The different types of phase separation and precipitate observed in many of the hanging drops might, therefore, have resulted from the unordered aggregation of protein molecules to exclude water from this hydrophobic surface. However, analysis of the crystal structure of BoNT/A revealed that the interface between the $H_C$ and $H_N$ domains is relatively small – the binding domain buries only ~400Å$^2$ of the translocation domain (Lacy and Stevens 1999). Close inspection of the surface of $LH_N/A$ (Figure 36) reveals that the region of $H_N$ that is normally in contact with the $H_C$ domain contains both polar and hydrophobic residues (in blue and red, respectively). It does not appear to have a high density of hydrophobic groups, which suggests that aggregation to exclude solvent molecules from this region would probably not occur.

**Figure 36 - Surface Diagram of the LC and $H_N$ Domains of BoNT/A**



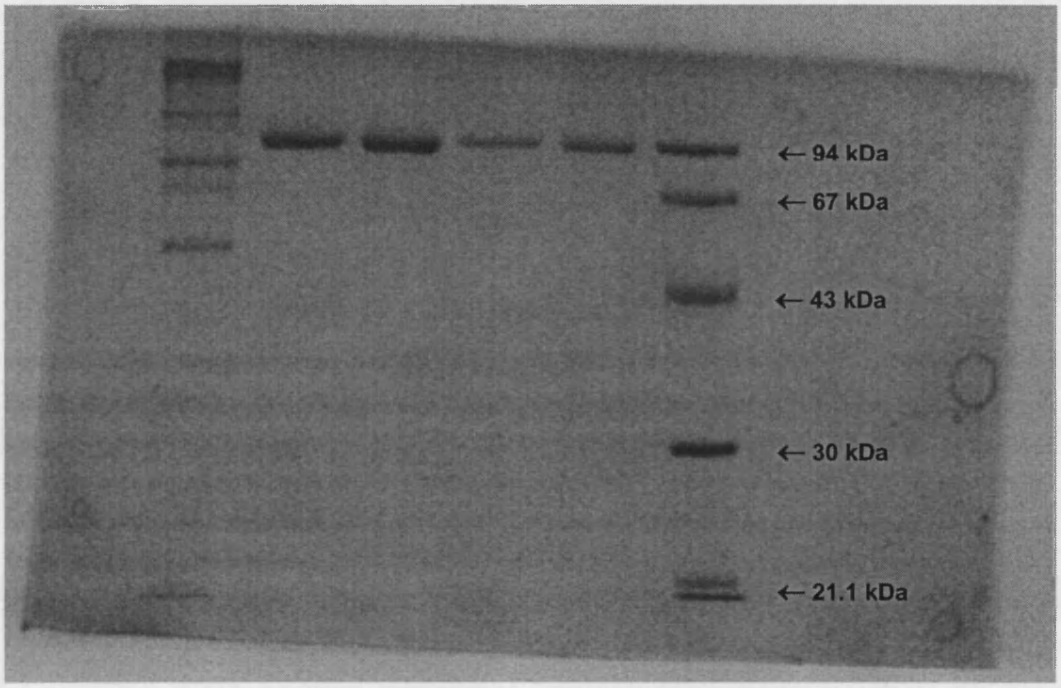

This figure was generated from the crystal structure of the BoNT/A holotoxin (PDB code 3BTA) using the program GRASP (Nicholls et al. 1991). Only residues of the LC and $H_N$ domains are displayed. The surface of the molecule is coloured according to the nature of the amino acids – hydrophobic residues red and polar residues blue. The region of $H_N$ that makes contacts with the $H_C$ domain (highlighted) contains both hydrophobic and polar residues. Therefore, the likelihood that $LH_N/A$ molecules aggregate in order to exclude solvent molecules from the region that normally binds $H_C$ (thus hindering crystallisation) is low.

#### 4.3.1.3 SDS-PAGE

The possibility that crystallisation trials were unsuccessful because the protein stocks had been stored at -20°C for a prolonged period of time was investigated. The purity of the protein samples had been tested by SDS-PAGE before despatch from CAMR (results not shown). Upon receipt from CAMR, the protein stocks had been divided into 100µl aliquots to minimise the need for repeated freeze-thaw, but it was suggested that the protein might not be stable at -20°C for more than a few months. After 12 months, the proteins were analysed again by "native" SDS-PAGE (i.e. with no dithiothreitol in the sample buffer) to ensure that the fragments had not degraded during storage at -20°C. The two ~50kDa domains comprising the $LH_N/A$ fragment are tethered solely by a single disulphide bond, which would be reduced by dithiothreitol. Native SDS-PAGE would therefore determine whether or not the disulphide bond linking the domains of $LH_N/A$ and $LH_N/A(H^{227}Y)$ had been reduced during storage.

The results (Figure 37) clearly show that the $LH_N/A$ and $LH_N/A(H^{227}Y)$ fragments had not degraded during storage at -20°C over 12 months. Based on analysis of their primary structures, the molecular weights of $LH_N/A$ and $LH_N/A(H^{227}Y)$ were calculated as 100052.6Da and 100078.6Da, respectively. Both of these fragments ran slightly higher on the gel than the 94kDa marker, in accordance with their molecular weights. This was another indication that the fragments had not been degraded during storage.

**Figure 37 - SDS-PAGE of $LH_N/A$ and $LH_N/A(H^{227}Y)$**




Photograph of the gel analysing the stocks of $LH_N/A$ and $LH_N/A(H^{227}Y)$ after 12 months stored at -20°C. High and Low Molecular Weight Markers were loaded into lanes 1 and 6, respectively. Both $LH_N/A(H^{227}Y)$, lanes 2 and 3, and $LH_N/A$, lanes 4 and 5, ran as single bands on the gel, indicating that the toxin fragments were not degraded during long-term storage at -20°C. The calculated molecular weight of each fragment is ~100kDa and both proteins ran on the gel slightly higher than the 94kDa marker.

### 4.3.2 Circular Dichroism Spectroscopy

Circular dichroism (CD) spectroscopy measures differences in the absorption of left-handed versus right-handed circularly-polarised light that arise due to structural asymmetry within the molecule under investigation (Kelly and Price 1997). Both right- and left-handed circularly-polarised light obey Beer's law and CD is defined as the difference in extinction coefficients:

$$\Delta A(\lambda) = A_L\lambda - A_R\lambda$$
$$= [\varepsilon_L(\lambda) - \varepsilon_R(\lambda)]$$
$$= \varepsilon lc$$

where: A is the optical absorbance of the molecule

ε is the exctinction coefficient of the molecule

λ is the wavelength of the incident light

c is the concentration of the molecule

(NB: the subscripts indicate the "handedness" of the light).

A CD spectrum is recorded by measuring dichroism as a function of wavelength (Kelly and Price 1997), with CD commonly presented in terms of ellipticity.

The asymmetry of the peptide bond makes proteins suitable for analysis by CD spectroscopy. However, the technique can only be used to determine the gross secondary and tertiary structural features of a protein (i.e. it can reveal the amount of protein involved in α-helix and β-sheet conformations, but cannot assign specific residues to these structural elements). The composition of secondary structural elements is determined by recording CD spectra in the "far-UV" spectral region (250-180nm) where the peptide bond acts as a chromophore. A CD signal is recorded when the peptide bond is located in a regular, folded environment. α-helix, β-sheet and random coil structures give rise to a characteristic shape and magnitude of CD spectrum. CD spectra recorded in the "near-UV" spectral region (250-350nm) provide information on the tertiary structure of a protein. At these wavelengths, aromatic residues and disulphide bonds act as chromophores and CD signals indicate that the protein is folded into a well-defined structure.

The structure of BoNT/A in aqueous solution is reported to differ from the crystal structure in terms of both secondary and quarternary structure (Cai and Singh 2001). Most notably, the crystal structure of BoNT/A (Lacy et al. 1998) reveals that ~17% of the structure is in a β-sheet conformation, whereas data obtained by infrared spectroscopy suggest that ~47% of the residues are in a β-sheet structure. The proportion of residues estimated to occupy β-sheet structures by CD spectroscopy is 44% (Singh and DasGupta 1989b) and 36% by primary structure analysis (Lebeda and Olson 1994). Clearly, there is some discrepancy regarding the structure of the toxin in aqueous solution and whether or not the crystal structure accurately reflects the native conformation of the protein.

In the absence of crystals of $LH_N/A$ and $LH_N/A(H^{227}Y)$, CD spectroscopy was employed to glean information on their structures. Far-UV CD spectra were recorded (across the range 185-250nm) to provide information on the secondary structure of

each fragment. Based upon the crystal structure of BoNT/A (Lacy et al. 1998) it was expected that the fragment would contain a large proportion of α-helical structure. It was also hoped that crystals would eventually be forthcoming to allow comparison of the solution and crystal structures of $LH_N/A$ and $LH_N/A(H^{227}Y)$.

#### 4.3.2.1 Comparison of CD Spectrum Analysis Methods

It has been proposed that the best analysis is obtained by comparing the results of these three methods (Sreerama and Woody 2000). However, if the analyses provided by the different methods vary widely, the estimates of secondary structure should not be considered reliable. The three methods used for CD spectrum analysis (CDSSTR, CONTIN and SELCON) were based upon different deconvolution algorithms. Previous comparison of these programs (Sreerama and Woody 2000) showed that they perform comparably in the deconvolution of many CD spectra. Indeed, they provided similar estimates of the secondary structural composition of $LH_N/A$ and $LH_N/A(H^{227}Y)$. It can be seen, from the fit of the experimental data to the reconstructed data (Appendix 4), that CDSSTR provided the closest match in all the analyses performed. However, the method that provides the closest correlation between experimental and reconstructed data does not always give the most accurate estimate of secondary structure. A better measure of accuracy is to use several deconvolution methods and look at how closely matched the results from all the analyses are. Ultimately, due to the high level of agreement between the analyses provided by the three methods, the estimated secondary structure for each fragment can be considered reliable.

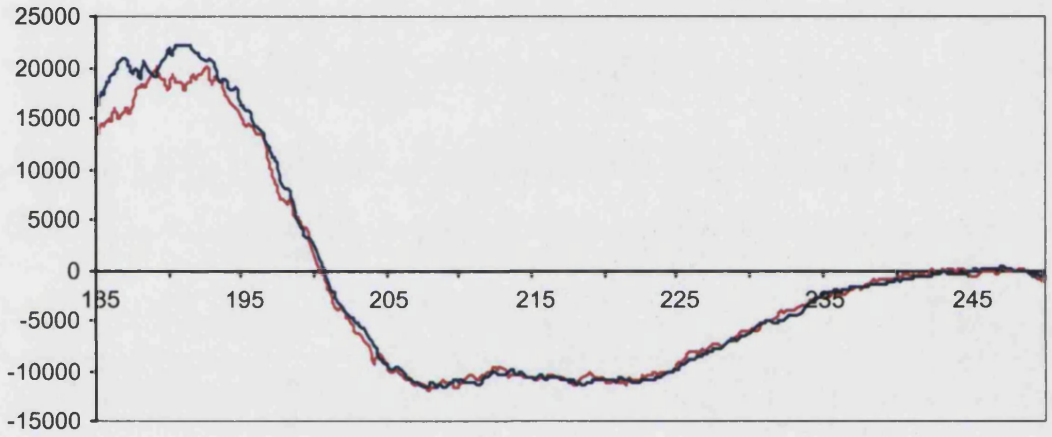
CD spectrum analysis is improved by using a protein reference that includes the largest possible representation of spectral features and secondary structural combinations (Sreerama and Woody 2000). Two protein reference sets spanning the wavelength range 185-240nm were tested, one containing data on five denatured proteins (Reference Set 6) and one with no denatured proteins (Reference Set 3). The raw data from the spectral analysis (Appendix 4) show that both reference sets provide similar estimates of secondary structural composition using either reference set.

CDSSTR (Compton and Johnson 1986; Manavalan and Johnson 1987; Sreerama and Woody 2000) uses a fixed number of proteins from the reference set (termed the "basis set"). Characteristics not reflected by the test protein are removed from the basis set and secondary structure is assigned. This method is believed to produce the most accurate analysis results. With CONTIN (Provencher and Glockner 1981; Van Stokkum et al. 1990), the CD spectrum of the unknown protein is fitted to a

set of "standard" spectra. SELCON3 (Sreerama et al. 1999; Sreerama and Woody 1993) arranges reference proteins in the database in order of increasing rmsd from the CD spectrum of the protein of interest (those that are most dissimilar are deleted). The structure of the unknown protein is estimated and included in the database. Spectra in the database are then related to a known mixture of secondary structures and used to analyse the conformation of the protein under investigation. The solution replaces the initial estimate and the process is repeated until self-consistency is attained.

#### 4.3.2.2 Structures of $LH_N/A$ and $LH_N/A(H^{227}Y)$

**Figure 38 - CD Spectra Recorded for $LH_N/A$ and $LH_N/A(H^{227}Y)$ at pH 7.0**




The CD spectra recorded for $LH_N/A$ (in red) and $LH_N/A(H^{227}Y)$ (in blue) at pH 7.0 show that the two toxin fragments have almost identical structures at the secondary level. The shapes of the spectra reveal that the proteins comprise some α-helical and some β-strand elements and deconvolution of the recorded data estimated the composition of $LH_N/A$ to be approximately 33% α-helix, 18% β-strand and 31% random coil.

The samples of $LH_N/A$ and $LH_N/A(H^{227}Y)$ used to record CD spectra at pH 7.0 had protein concentrations of 0.106mg/ml and 0.186mg/ml, respectively. Molar extinction coefficients at 280nm were calculated as 94549 $M^{-1}cm^{-1}$ for $LH_N/A$ and 95820 $M^{-1}cm^{-1}$ for $LH_N/A(H^{227}Y)$ using amino acid sequence data for the toxin fragments (http://ca.expasy.org/tools/#protparam.html; Gill and von Hippel 1989) and the mean residue weights for $LH_N/A$ and $LH_N/A(H^{227}Y)$, were calculated as 114.608 and 114.638, respectively. The far-UV CD spectra recorded for $LH_N/A$ and $LH_N/A(H_{227}Y)$ at pH 7.0 (Figure 38) confirm that the toxin fragments are almost identical in terms of their structures at the secondary level. Analysis of these spectra (Appendix 4) revealed that $LH_N/A$ is composed of approximately 33% α-helix, 18% β-strand, 19% β-turn and 31%

random coil, and that $LH_N/A(H^{227}Y)$ comprises ~34% α-helix, ~19% β-strand, ~18% β-turn and ~30% random coil.

It was expected that the two fragments would have nearly identical secondary structures because there is only one amino acid different between them. Whilst the mutation renders the $LH_N/A(H^{227}Y)$ fragment catalytically inactive, the side chains of histidine and tyrosine are of comparable size, so it is unlikely that the mutation will alter the tertiary structure. The imidazole ring of histidine acts as a nucleophile and is a strong base at neutral pH, whereas tyrosine is an aromatic residue that is nonpolar and takes part in electrophilic substitutions. Thus, the $H^{227}Y$ mutation leads to inactivity of $LH_N/A$ and BoNT/A by altering the chemical composition of the active site and not by causing a conformational change that prevents the reaction from taking place.

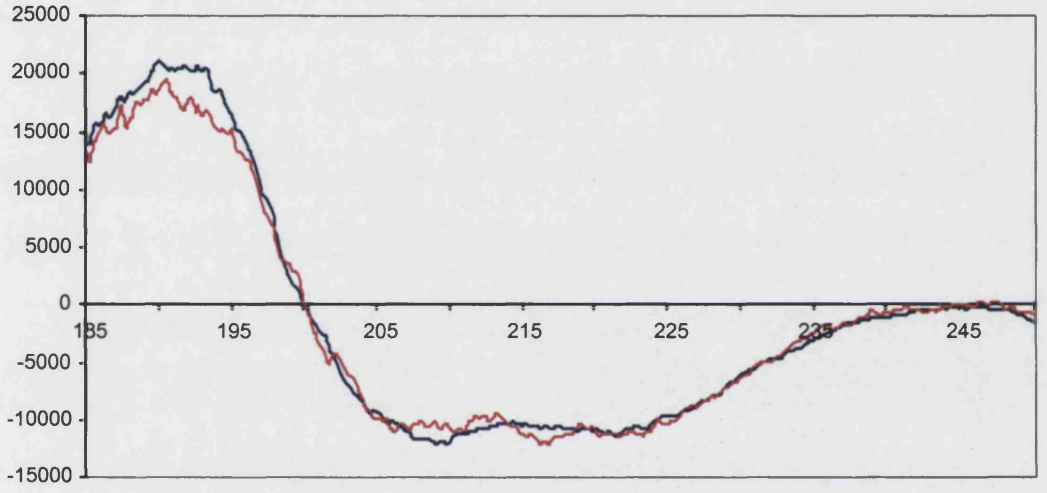
To my knowledge, no previous structural analyses have been performed on the $LH_N/A$ fragment. However, CD experiments performed on the BoNT/A-LC alone revealed that it is 22% α-helical and it was concluded that the secondary structures of the light and heavy chains of BoNT/A are not altered upon their separation (Singh and DasGupta 1989a). Crystallography revealed the CNT $H_N$ domain to be mainly α-helical, with the main feature being two long (105Å), amphipathic helices running anti-parallel to, and twisting around, each other. The LC domain, on the other hand, comprises a mixture of α-helical and β-strand structural elements. Using secondary structural information from the PDB file for the BoNT/A holotoxin (PDB code 3BTA), the proportion of residues in the LC and $H_N$ domains of the crystal structure was calculated as 39.% helix and 7% strand. This simple analysis reveals that the proportion of residues adopting helical conformation is comparable between the crystal structure and the solution structure. However, the CD spectroscopy studies indicate that there is a greater proportion of β-strand in the solution structure of $LH_N/A$ than in the crystal structure.

#### 4.3.2.3 Structure of $LH_N/A(H^{227}Y)$ at Neutral and Endosomal pH

It has long been known that BoNT/A forms ion-conducting channels, a function assigned to the toxin heavy chain. The formation of channels in planar lipid membranes was observed to be pH-dependent, with the maximum rate of channel formation achieved at pH 6.1 (Donovan and Middlebrook 1986). Further experiments performed in lipid bilayers showed that channel formation requires low pH (4.5-5.5) on one side of the membrane and neutral pH on the other, as had previously been seen for diphtheria toxin and TeNT (Blaustein et al. 1987). BoNT/A channels formed in phospholipids vesicles have been observed to extend across the vesicle wall and appear to arise

from the interaction of protein molecules on both sides of the channel (Schmid et al. 1993). However, the predicted channel-forming regions of BoNT/B-$H_N$ do not adopt a helical conformation in the crystal structure (Swaminathan and Eswaramoorthy 2000b), which suggests that the BoNT heavy chain is subject to conformational changes at low pH that are conducive with channel formation. The tertiary structure of BoNT/A-LC was recently observed to undergo conformational changes at acidic pH (Li and Singh 2000b). Therefore, far-UV CD spectra were recorded to determine whether or not there is a change in the secondary structural composition of $LH_N/A(H^{227}Y)$ at endosomal pH.

**Figure 39 - CD Spectra Recorded for $LH_N/A(H^{227}Y)$ at pH 5.0 and 7.0**




The CD spectra reveal that there is no significant change in the secondary structure of $LH_N/A(H^{227}Y)$ at low pH (in red) compared with its structure at neutral pH (in blue). This was an unexpected result because both the LC and $H_N$ domains are known to undergo conformational changes in acidic environments.

The protein concentrations of the $LH_N/A(H^{227}Y)$ samples at pH 7.0 and 5.0 used in later experiments were 0.289mg/ml and 0.133mg/ml, respectively. Analysis of the spectra (Figure 39; Appendix 4) revealed that the secondary structure of $LH_N/A(H^{227}Y)$ at pH 5.0 is no different to its structure at pH 7.0.

There are two possible explanations for the results obtained. Firstly, it has been shown that the heavy chain alone of BoNT/A or BoNT/A can form functional ion channels (Blaustein et al. 1987). Thus, it might be postulated that pore-formation requires the presence of the $H_C$ domain, perhaps to contribute some structural stability for channel formation or additional functional groups for ion conductance. As $LH_N/A(H^{227}Y)$ is devoid of the binding domain, this influence over channel-formation is

absent and might explain why no structural changes were observed at low pH. Secondly, previous studies on pore formation by BoNTs were performed in systems that comprised lipid bilayers or artificial membranes and it was shown a requirement for channel formation is low pH on the "endosomal" side of the membrane and neutral pH on the "cytosolic" side. The experimental system used to study the secondary structure of $LH_N/A(H^{227}Y)$, however, contained no lipids or bilayers. It is likely that the lack of a pH gradient and the absence of an artificial membrane system prevented the expected structural changes occurring.

## 4.4 SUMMARY OF RESULTS

Despite setting up several hundred hanging drops, crystallisation experiments were unsuccessful and no crystals of either $LH_N/A$ or $LH_N/A(H^{227}Y)$ were grown during the course of this investigation. Several commonly used precipitants were tested at a variety of concentrations in conjunction with several buffers over the pH range 4.6 to 9.0. Although two unusual forms of phase separation, microcrystals and shiny crystalloid structures, were observed, attempts to optimise the conditions producing these phenomena to yield real crystals of diffraction quality failed. The possibility of crystallisation under oil was ruled out because the most promising results were obtained with mother liquors comprising alcohols, such as MPD. These compounds are soluble in paraffin oil, rendering these conditions unsuitable for crystallisation by this method. The technique of streak seeding was reported to be essential for the growth of BoNT/A crystals (Swaminathan and Eswaramoorthy 2000a) but did not aid crystallisation of $LH_N/A$ or $LH_N/A(H^{227}Y)$. Analysis of the fragments by non-reducing SDS-PAGE confirmed that the proteins had not degraded during storage at -20°C.

Far-UV CD spectroscopy revealed that there were no significant structural differences between $LH_N/A$ and $LH_N/A(H^{227}Y)$ in terms of their secondary structures. The secondary structural composition of both fragments is ~33% α-helix, ~18% β-strand and ~31% random coil. The high degree of similarity between the structures of $LH_N/A$ and $LH_N/A(H^{227}Y)$ was expected because there is only one amino acid different between the two toxin fragments. In light of reported evidence for conformational changes within the $H_N$ and LC domains of BoNT/A at low pH, it was expected that differences in the structure of $LH_N/A(H^{227}Y)$ at endosomal pH would be observed compared to its structure at neutral pH, but no such differences were observed via CD spectroscopy.

# CHAPTER FIVE
# CONCLUSION

As described in Chapter One, the $LH_N$/A-ECL conjugate is of potential use as a novel biopharmaceutical product in the treatment of chronic pain syndromes. The biology underlying this therapeutic application is founded upon the ability of botulinum neurotoxin A to ablate neurotransmission in intoxicated cells for up to 6 months and the specificity of *Erythrina cristagalli* lectin for galactose-containing glycoconjugates that are selectively expressed on nociceptive afferents. The ultimate aim of the project was to increase our understanding of the biology of the protein components of the $LH_N$/A-ECL conjugate by studying their three dimensional structures.

## 5.1 CRYSTAL STRUCTURE OF *ERYTHRINA CRISTAGALLI* LECTIN

Native and recombinant forms of ECL were crystallised by the hanging drop vapour diffusion method in five crystal forms under a variety of conditions. Both forms of ECL were successfully co-crystallised with lactose and *rec*ECL was also crystallised in complex with galactose and N-acetylgalactosamine, although the resulting crystals were not suitable for diffraction experiments. A total of six crystal forms were analysed and the three-dimensional structures of *n*ECL and *rec*ECL determined to atomic resolution.

The tertiary structure of *Erythrina cristagalli* lectin is homologous to the known structures of other legume lectins. ECL protomers adopt the jelly-roll (legume lectin) fold and associate together back-to-back to form dimers via the handshake motif. Each protomer contains one manganese ion and one calcium ion, both of which are located close to the carbohydrate-combining site. Furthermore, the structure of the oligosaccharide bound to $Asn^{113}$ on native ECL was modelled. There are no significant differences between the crystal structures of native and recombinant ECL, confirming that the recombinantly over-expressed form has a native-like structure. As the conjugate is only manufactured using recombinant ECL, it is important to demonstrate that this represents the native form of the lectin in terms of both its structure and function. Equivalence of biological activity has already been proved (Stancombe et al. 2003) and this study confirms structural homology. The only difference between *n*ECL and *rec*ECL is the absence of glycosylation from the recombinant form – a consequence of using *Escherichia coli* as an expression host. Direct comparison of the

structures of *n*ECL and *rec*ECL revealed that the presence of bound carbohydrate at $Asn^{113}$ affects neither the structure nor carbohydrate-binding activity of the lectin.

Analysis of the structure of ECL in complex with lactose revealed the mode of carbohydrate binding to be similar to that observed in ECorL. Lactose makes a number of interactions with the lectin through the galactose moiety, including indirect hydrogen bonds mediated by structural water molecules and hydrophobic contacts. The key residues involved in lactose binding at the combining site are $Leu^{86}$, $Asp^{89}$, $Tyr^{106}$, $Gly^{107}$, $Phe^{131}$, $Asn^{133}$, $Ala^{218}$ and $Gln^{219}$. Superposition of the combining sites of native and recombinant forms of ECL revealed that both forms of the lectin bind carbohydrates in the same way, confirming that the recombinant form of ECL is native-like with respect to its mode of carbohydrate recognition as well as in terms of its three-dimensional structure. Furthermore, comparison of the combining sites of bound and unbound forms of ECL showed that there are no structural changes in the vicinity of the combining site upon lactose binding. This indicates that the lectin combining site is pre-formed within each protomer and suggests that lectins interact with their ligands via a "lock-and-key" mechanism.

Comparison of the structure of *n*ECL with that recently reported for native ECL (Svensson et al. 2002) revealed several differences in the amino acid sequences of the two proteins, although there were no significant differences in their structures at the tertiary or quarternary levels. One of the amino acid substitutions removes a potential N-linked glycosylation site from *n*ECL at position 17 (which is aspartic acid in *n*ECL, but asparagine in *Sven*ECL). It was postulated that the presence of a heptasaccharide at $Asn^{17}$ in both ECorL and ECL forced these lectins to dimerise via the non-canonical handshake motif (Shaanan et al. 1991; Svensson et al. 2002). However, neither *n*ECL nor *rec*ECL is glycosylated at this position, yet both form ECorL-type dimers, suggesting that glycosylation does not influence the quarternary association of these lectins. Taken together, these observations suggest that factors intrinsic to the primary structures of the legume lectins determine their mode of oligomerisation.

## 5.2 STRUCTURAL STUDIES ON BOTULINUM NEUROTOXIN A

No crystals of $LH_N/A$ or $LH_N/A(H^{227}Y)$ were grown, which meant that an alternative technique to X-ray crystallography had to be sought in order to glean information on the structures of these proteins. The method used to analyse $LH_N/A$ and $LH_N/A(H^{227}Y)$ in terms of their secondary structures was far-UV circular dichroism spectroscopy. At 250-180nm the peptide bond acts as a chromophore when it is arranged in a regular repeating structure, such as an α-helix or β-sheet. Each type of

secondary structural element produces a characteristic spectrum in the far-UV wavelength range. Using a deconvolution algorithm, the spectrum recorded for a protein of unknown structure can be analysed and the percentage of amino acid residues involved in each type of structural element estimated. One disadvantage of this technique is that the identities of the residues involved in, for example, an α-helix remain elusive.

Using CD spectroscopy, the structures of $LH_N/A$ and $LH_N/A(H^{227}Y)$ in aqueous solution were directly compared and found to be highly similar. Overall, approximately 33% of amino acid residues were arranged in helical structures, 18% in β-strand elements and 31% in random coil conformations. Furthermore, despite evidence to suggest a structural rearrangement in the light and heavy chains of BoNT/A under the influence of low pH, the structure of $LH_N/A(H^{227}Y)$ at pH 5.0 was not significantly different to its structure at neutral pH.

## 5.3 SCOPE OF THE PROJECT

To my knowledge, the data generated during the course of this investigation represent the first studies into the structures of $LH_N/A$ and $LH_N/A(H^{227}Y)$ and the first description of the crystal structure of recombinantly expressed ECL. Furthermore, the structure of *rec*ECL has been shown to be identical to the structure of *n*ECL, despite the absence of glycosylation, and for the first time the structure of the oligosaccharide N-linked to $Asn^{113}$ of native ECL has been described.

### 5.3.1 Crystal Structures of ECL

By comparing the crystal structures of *n*ECL (glycosylated at $Asn^{113}$) and *rec*ECL (not glycosylated) with the reported structure of ECL (glycosylated at $Asn^{17}$ and $Asn^{113}$) we have shown that the presence of heptasaccharide bound to $Asn^{17}$ does not influence the quarternary structure of the lectin. This contradicts previous suggestions that ECL and ECorL adopt non-canonical dimer structures because the carbohydrate bound at this position prevents canonical dimerisation and supports data suggesting that information controlling the mode of quarternary association in legume lectins is encoded within the amino acid sequence of each protomer.

It was important to demonstrate the equivalence of native and recombinant ECL since the $LH_N/A$-ECL conjugate is prepared using recombinantly expressed ECL. The biological activities of both forms of the lectin *in vitro* have been shown to be identical (Stancombe et al, 2003) and the results of this investigation have confirmed that there is no difference between *n*ECL and *rec*ECL either in terms of their three-

dimensional structure or their mode of carbohydrate binding. Thus, it can be concluded that *rec*ECL represents a native-like form of *Erythrina cristagalli* lectin.

### 5.3.2 Structures of BoNT/A Fragments

Far-UV CD studies allowed the solution structures of $LH_N/A$ and $LH_N/A(H^{227}Y)$ to be directly compared. The structures of these two fragments are nearly identical but have not previously been described. The failure of crystallisation trials raised several questions about the structures of these toxin fragments, in particular whether or not they are stable in solution. SDS-PAGE and CD spectroscopy have revealed that they are stable at -20°C for at least 12 months because a secondary structural profile could be determined and they ran as single bands on a gel. This is encouraging, since the use of these fragments in a biopharmaceutical might require that they should be stable during several months' storage. However, little information could be deduced regarding their biological activities from the CD data and future work on $LH_N/A$ and $LH_N/A(H^{227}Y)$ should focus on the determination of their three-dimensional structures.

## 5.4 FUTURE DIRECTIONS

### 5.4.1 BoNT/A Fragments

Further crystallisation experiments should be undertaken, including the use of substrate analogues and active-site inhibitors to try to yield crystals of $LH_N/A$ and $LH_N/A(H^{227}Y)$ suitable for diffraction experiments. The addition of an inhibitor or substrate analogue can provide extra stability to the protein and aid crystallisation. In addition, a wider range of alcohols, additives and environmental variables (for example, temperature) could be tested to try to crystallise these fragments. Synthetic peptides that act as inhibitors and substrate analogues have already been identified (Hallis et al. 1996; Schmidt et al. 1998), but none of these compounds were tested during the course of this project. The possibility of using detergents as additives was considered as they have been shown to be useful in the crystallisation of membrane proteins (Abramson and Iwata 1999). However, this idea was based upon the assumption that $LH_N/A$ has a large hydrophobic region where the $H_C$ domain would normally exist and crystallographic data reveal suggest this is not the case (Lacy et al. 1998).

In the absence of a crystal structure, further far-UV CD studies involving the presence of SDS micelles or lipids could also be performed to try to observe conformational changes induced in $LH_N/A$ and $LH_N/A(H^{227}Y)$ at low pH. In parallel, near-UV CD spectra could be recorded to characterise and look for changes in the

tertiary structures of these toxin fragments in similar experiments to those performed in this investigation.

### 5.4.2 *Erythrina cristagalli* Lectin

In this investigation, *n*ECL was not successfully crystallised in the absence of lactose and only one crystal form of *n*ECL was analysed compared to five different forms of *rec*ECL. Determination of the structures of both forms of ECL in complex with a variety of carbohydrates is desired to gain a more comprehensive understanding of carbohydrate recognition by ECL. In particular, it would be of interest to analyse the interactions of ECL with larger oligosaccharides and glycoconjugates that more closely represent the natural ligands of the lectin to assess the involvement of dimerisation and cross-linking in its biological function.Since the lectin combining site is preformed and each protomer contains the structural components required to be fully functional in carbohydrate recognition, it is not known why ECL exists as a dimer. Although ECL has never been isolated in monomeric form, it might be possible to retarget $LH_N/A$ to target cells using only one protomer. If the lectin can be isolated in monomeric form, it will be interesting to compare the activities of a monomeric ECL-$LH_N/A$ conjugate with the existing conjugate to further assess the effect of dimerisation on ECL activity.

### 5.4.3 $LH_N/A$-ECL Conjugate

Combining data on the crystal structures of $LH_N/A$ and ECL might enable the three-dimensional structure of the $LH_N/A$-ECL conjugate to be modelled. However, any such model could not be relied upon to be accurate because the exact structure of the conjugate is unknown. Crystallisation and structural analysis of $LH_N/A$-ECL itself (both unliganded and in complex with its natural ligands) would provide the most useful information as the action of this conjugate on dorsal root ganglia compared to the anatomically adjacent spinal cord neurons is the basis of its potential use as a pharmaceutical. However, at the moment, this is not feasible for several reasons:

Firstly, the conjugation method produces a mixture of products because the proteins are non-specifically linked through a primary amine on the surface of the $LH_N/A$ moiety. It is not known exactly where the lectin and toxin fragment are joined or how many different products are formed. Therefore, the exact point of chemical linkage cannot be predicted and several different linkages might be formed during each conjugation reaction. Furthermore, the stoichiometry of the conjugate is believed to be two ECL dimers per unit of $LH_N/A$. However, this represents the average mass of all reaction products. Some species may comprise different combinations of dimeric ECL and $LH_N/A$. At present, it is not known which of the products are active *in vitro* and *in*

*vivo*, although it is hoped that further purification of the conjugated material will shed light on the nature of the true potential biopharmaceutical. Improving homogeneity will hopefully increase the biological activity of the conjugate as less active and inactive species will be removed from the purified material.

The heterogeneity of the final product is expected to seriously hinder crystallisation attempts. Furthermore, if the conjugate comprises two ECL dimers and one $LH_N/A$ moiety, the molecular mass would be close to 205kDa, which might also make crystallisation difficult. Larger molecules do not tend to pack together very tightly, which can reduce the likelihood of forming stable nuclei that would induce crystal growth.

# BIBLIOGRAPHY

Collaborative Computational Project Number 4. 1994. The CCP4 Suite: Programs for Protein Crystallography *Acta Crystallogr D* 50:760-763.

Abramson J. and Iwata S. 1999. Crystallization of Membrane Proteins. In: Bergfors T.M. ed *Protein Crystallization: Techniques, Strategies, and Tips*, International University Line, pp 197-210

Adams P.D., Pannu N.S., Read R.J. and Brunger A.T. 1997. Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement *Proc Natl Acad Sci USA* 94:5018-5023.

Adar R., Moreno E., Streicher H., Karlsson K.A., Angstrom J. and Sharon N. 1998. Structural features of the combining site region of Erythrina corallodendron lectin: Role of tryptophan 135 *Prot Sci* 7:52-63.

Adler M., Nicholson J.D. and Hackley B.E.J. 1998. Efficacy of a novel metalloprotease inhibitor on botulinum neurotoxin B activity *FEBS Lett* 429:234-238.

Ahnert-Hilger G. and Bigalke H. 1995. Molecular aspects of tetanus and botulinum neurotoxin poisoning *Prog Neurobiol* 46:83-96.

Allison M.C., Howatson A.G., Torrance C.J., Lee F.D. and Russell R.I. 1992. Gastrointestinal damage associated with the use of nonsteroidal anti-inflammatory drugs *N Engl J Med* 327:749-754.

Amemiya Y. 1997. X-Ray Storage-Phosphor Imaging-Plate Detectors: High-Sensitivity X-Ray Area Detector *Methods Enzymol* 276:233-243.

Aoki K.R. 2003. Evidence for Antinociceptive Activity of Botulinum Toxin Type A in Pain Management *Headache* 43:S9-S15.

Ashburn M.S. and Staats P.S. 1999. Management of chronic pain *Lancet* 353:1865-1869.

Ashford D.A., Dwek R.A., Rademacher T.W., Lis H. and Sharon N. 1991. The glycosylation of glycoprotein lectins. Intra- and inter- genus variation in N-linked oligosaccharide expression *Carbohydr Res* 213:215-227.

Ashford D.A. and Platt F. 1999. Protein glycosylation. In: Higgins S.J., Hames B.D. eds., *Post-translational Processing: A Practical Approach*, Oxford University Press

Ashton A., Li Y., Doussau F., Weller U., Dougan G., Poulain B. and Dolly J.O. 1995. Tetanus Toxin Inhibits Neuroexocytosis Even When Its $Zn^{2+}$-dependent Protease Activity Is Removed *J Biol Chem* 270:31386-31390.

Ashwell G. and Harford J. 1982. Carbohydrate-specific receptors of the liver *Annu Rev Biochem* 51:531-554.

Aubery M., Reynier M., Lopez M., Ogier-Denis E., Font J. and Bardin F. 1990. WGA binding to the surface of two autologous human melanoma cell lines: different expression of sialyl and N-acetylglucosaminyl residues *Cell Biol Int Rep* 14:275-286.

Aversa F., Tabillo A., Terenzi A., Velardi A., Falzetti F., Giannoni C., Iacucci R., Zel T., Martelli M.P., Gambelunghe C., Rossetti M., Caputo P., Latini P., Aristel C., Raymondi C., Reisner Y. and Martelli M.F. 1994. Successful Engraftment of T-Cell-Depleted Haploidentical "Three-Loci" Incompatible Transplants in Leukaemia Patients by Addition of Human Granulocyte Colony-Stimulating Factor-Mobilized Peripheral Blood Progenitor Cells to Bone Marrow Inoculum *Blood* 84:3948-3955.

Bailey S. 1994. The CCP4 Suite: Programs for Protein Crystallography *Acta Crystallogr D* 50:760-763.

Banerjee R., Mande S.C., Ganesh V., Das K., Dhanaraj V., Mahanta S.K., Suguna K., Surolia A. and Vijayan M. 1994. Crystal structure of peanut lectin, a protein with an unusual quarternary structure *Proc Natl Acad Sci USA* 91:227-231.

Banerjee S., Robson P., Soutter W.P. and Foster C.S. 1995. Modulated expression of glycoprotein oligosaccharides identifies phenotypic differentiation in squamous carcinomas of the human cervix *Hum Pathol* 26:1005-1013.

Bergelson L.D., Bukrinskaya A.G., Prokazova N.V., Shaposhnikova G.I., Kocharov S.L., Shevchenko V.P., Kornilaeva G.V. and Fomina-Ageeva E.V. 1982. Role of gangliosides in reception of influenza virus *Eur J Biochem* 128:467-474.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N. and Bourne P.E. 2000. The Protein Databank *Nucl Acid Res* 28:235-242.

Bevilacqua M.P., Stengelin S., Gimbrone M.A.J. and Seed B. 1989. Endothelial leukocyte adhesion molecule 1: an inducible receptor for neutrophils related to complement regulatory proteins and lectins *Science* 243:1160-1165.

Binz T., Bade S., Rummel A., Kollewe A. and Alves J. 2002. Arg362 and Tyr365 of the Botulinum Neurotoxin Type A Light Chain are Involved in Transition State Stabilization *Biochemistry* 41:1717-1723.

Blaustein R.O., Germann W.J., Finkelstein A. and DasGupta B.R. 1987. The N-terminal half of the heavy chain of botulinum type A neurotoxin forms channels in planar phoshplipid bilayers *FEBS Lett* 226:115-120.

Blundell T.L. and Johnson L.N. 1976. ,*Protein Crystallography*, Academic Press.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Michoud K., O'Donovan C., Phan L., Pilbout S. and Schneider M. 2003. The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003 *Nucl Acid Res* 31:365-370.

Bohlool B.B. and Schmidt E.L. 1974. Lectins: a possible basis for specificity in the Rhizobium-legume root nodule symbiosis *Science* 185:269-271.

Bouckaert J., Loris R., Poortmans F. and Wyns L. 1995. Crystallographic structure of metal-free concanavalin A at 2.5A resolution *Proteins: Struct Funct Genet* 23:510-524.

Bouckaert J., Poortmans F., Wyns L. and Loris R. 1996. Sequential Structural Changes upon Zinc and Calcium Binding to Metal-free Concanavalin A *J Biol Chem* 271:16144-16150.

Bourne Y., Bolgiano B., Liao D.I., Strecker G., Cantau P., Herzberg O., Feizi T. and Cambillau C. 1994. Cross-linking of mammalian hepatic lectin (galectin-1) by complex biantennary saccharides *Nat Struct Biol* 1:863-870.

Brunger A.T. 1992. Free R value: a novel statistical quantity for assessing the accuracy of crystal structures *Nature* 355:472-474.

Brunger A.T., Adams P.D., Clore G.M., DeLano W.L., Gros P., Grosse-Kustleve R.W., Jiang J.-S., Kuszewski J., Nilges M., Pannu N.S., Read R.J., Rice L.M., Simonson T.

and Warren G.L. 1998. Crystallography and NMR System: A New Software Suite for Macromolecular Structure Determination *Acta Crystallogr D* 54:905-921.

Brunger A.T., Adams P.D. and Rice L.M. 1997. New applications of simulated annealing in X-ray crystallography and solution NMR *Structure* 5:325-326.

Brunger A.T., Krukowski K. and Erickson J.W. 1990. Slow-cooling protocols for crystallographic refinement by simulated annealing *Acta Crystallogr A* 46:585-593.

Brunger A.T., Kuriyan J. and Karplus M. 1987. Crystallographic R Factor Refinement by Molecular Dynamics *Science* 235:458-460.

Byrne M.P., Smith T.J., Montgomery V.A. and Smith L.A. 1998. Purification, Potency, and Efficacy of the Botulinum Neurotoxin Type A Binding Domain from Pichia pastoris as a Recombinant Vaccine Candidate *Infect Immun* 66:4817-4822.

Byrne M.P., Titball R.W., Holley J. and Smith L.A. 2000. Fermentation, Purification, and Efficacy of a Recombinant Vaccine Candidate against Botulinum Neurotoxin Type F from Pichia pastoris *Prot Expr Purif* 18:327-337.

Cai S., Sarkar H.K. and Singh B.R. 1999. Enhancement of the Endopeptidase Activity of Botulinum Neurotoxin by Its Associated Proteins and Dithiothreitol *Biochemistry* 38:6903-6910.

Cai S. and Singh B.R. 2001. A Correlation between Differential Structural Features and the Degree of Endopeptidase Activity of Type A Botulinum Neurotoxin in Aqueous Solution *Biochemistry* 40:4693-4702.

Chaddock J.A., Herbert M.H., Ling R.J., Alexander F.C., Fooks S.J., Revell D.F., Quinn C.P., Shone C.C. and Foster K.A. 2002. Expression and purification of catalytically active, non-toxic endopeptidase derivatives of Clostridium botulinum toxin type A *Protein Expr Purif* 25:219-28.

Chaddock J.A., Purkiss J.R., Duggan M.J., Quinn C.P., Shone C.C. and Foster K.A. 2000a. A conjugate composed of nerve growth factor coupled to a non-toxic derivative of Clostridium botulinum neurotoxin type A can inhibit neurotransmitter release in vitro *Growth Factors* 18:147-55.

Chaddock J.A., Purkiss J.R., Friis L.M., Broadbridge J.D., Duggan M.J., Fooks S.J., Shone C.C., Quinn C.P. and Foster K.A. 2000b. Inhibition of vesicular secretion in both

neuronal and nonneuronal cells by a retargeted endopeptidase derivative of Clostridium botulinum neurotoxin type A *Infect Immun* 68:2587-93.

Chayen N.E. 1997. The role of oil in macromolecular crystallization *Structure* 5:1269-1274.

Chen F., Kuziemko G.M. and Stevens R.C. 1998. Biophysical Characterization of the Stability of the 150-Kilodalton Botulinum Toxin, the Non-toxic Component, and the 900-Kilodalton Botulinum Toxin Complex Species *Infect Immun* 66:2420-2425.

Chen H., Torchilin V. and Langer R. 1996. Lectin-bearing polymerized liposomes as potential oral vaccine carriers *Pharm Res* 13:1378-1383.

Clark M.A., Hirst B.H. and Jepson M.A. 2000. Lectin-mediated mucosal delivery of drugs and microparticles *Adv Drug Delivery Rev* 43:207-223.

Clayton J. and Middlebrook J.L. 2000. Vaccination of mice with DNA encoding a large fragment of botulinum neurotoxin serotype A *Vaccine* 18:1855-1862.

Clayton M.A., Clayton J.M., Brown D.R. and Middlebrook J.L. 1995. Protective Vaccination with a Recombinant Fragment of Clostridium botulinum Neurotoxin Serotype A Expressed from a Synthetic Gene in Escherichia coli *Infect Immun* 63:2738-2742.

Clive D.M. and Stoff J.S. 1984. Renal Syndromes Associated with Nonsteroidal Antiinflammatory Drugs *N Engl J Med* 310:563-572.

Compton L.A. and Johnson W.C.J. 1986. Analysis of protein circular dichroism spectra using a simple matrix multiplication *Anal Biochem* 155:155-167.

Dahms N.M., Lobel P. and Kornfeld S. 1989. Mannose 6-Phosphate Receptors and Lysosomal Enzyme Targeting *J Biol Chem* 164:12115-12118.

Dauter Z. 1997. Data Collection Strategy *Methods Enzymol* 276:326-344.

Dauter Z. 1999. Data-collection strategies *Acta Crystallogr D* 55:1703-1717.

Dazzo F.B. and Hubbell H.D. 1975. Cross-reactive antigens and lectins as determinants of symbiotic specificity in the Rhizobium-clover association *Appl Microbiol* 30:1017-1033.

Dekleva M.L. and DasGupta B.R. 1989. Nicking of single chain Clostridium botulinum type A neurotoxin by an endogenous protease *Biochim Biophys Acta* 162:767-772.

Delbaere L.T.J., Vandonselaar M., Prasad L., Quail J.W., Wilson K.S. and Dauter Z. 1993. Structures of the Lectin IV of Griffonia simplicifolia and its Complex with the Lewis b Human Blood Group Determinant at 2.0A Resolution *J Mol Biol* 230:950-965.

dePaiva A., Meunier F.A., Molgo J., Aoki K.R. and Dolly J.O. 1999. Functional repair of motor endplates after botulinum neurotoxin type A poisoning: Biphasic switch of synaptic activity between nerve sprouts and their parent terminals *Proc Natl Acad Sci USA* 96:3200-3205.

Derewenda Z., Yariv J., Helliwell J.R., Kalb (Gilboa) A.J., Dodson E.J., Papiz M.W., Wan T. and Campbell J. 1989. The structure of the saccharide binding site of concanavalin A *EMBO J* 8:2189-2193.

Dessen A., Gupta D., Sabesan S., Brewer C.F. and Sacchettini J.C. 1995. X-ray crystal structure of the soybean agglutinin cross-linked with a biantennary analog of the blood group I carbohydrate antigen *Biochemistry* 34:4933-4942.

Donaldson M.S. and Shuler M.L. 1999. The use of lectins to select subpopulations of insect cells *Biotechnol Bioeng* 64:616-619.

Donovan J.J. and Middlebrook J.L. 1986. Ion-Conducting Channels Produced by Botulinum Toxin in Planar Lipid Membranes *Biochemistry* 25:2872-2876.

Donovan J.J., Simon M.I., Draper R.K. and Montal M. 1981. Diphtheria toxin forms transmembrane channels in planar lipid bilayers *Proc Natl Acad Sci USA* 78:172-176.

Drenth J. 1999. *Principles of Protein X-Ray Crystallography* Second Edition, Springer.

Ducruix A. and Giege R. 1999. *Crystallization of Nucleic Acids and Proteins: A Practical Approach* Second Edition

Duggan M.J., Quinn C.P., Chaddock J.A., Purkiss J.R., Alexander F.C., Doward S., Fooks S.J., Friis L.M., Hall Y.H., Kirby E.R., Leeds N., Moulsdale H.J., Dickenson A., Green G.M., Rahman W., Suzuki R., Shone C.C. and Foster K.A. 2002. Inhibition of release of neurotransmitters from rat dorsal root ganglia by a novel conjugate of a Clostridium botulinum toxin A endopeptidase fragment and Erythrina cristagalli lectin *J Biol Chem* 277:34846-52.

Eleopra R., Tugnoli V., Rossetto O., Montecucco C. and De Grandis D. 1997. Botulinum neurotoxin serotype C: a novel effective botulinum toxin therapy in human *Neurosci Lett* 224:91-94.

Elgavish S. and Shaanan B. 1998. Structures of the Erythrina corallodendron lectin and of its complexes with mono- and disaccharides *J Mol Biol* 277:917-32.

Elliott A.M., Smith B.H., Penny K.I., Smith W.C. and Chambers W.A. 1999. The epidemiology of chronic pain in the community *Lancet* 354:1248-1252.

Emmerich C., Helliwell J.R., Redshaw M., Naismith J.H., Harrop S.J., Raftery J., Kalb (Gilboa) A.J., Yariv J., Dauter Z. and Wilson K.S. 1994. High-Resolution Structures of Single-Metal-Substituted Concanavalin A: the Co,Ca-Protein at 1.6A and the Ni,Ca-Protein at 2.0A *Acta Crystallogr D* 50:749-756.

Emsley P., Fotinou C., Black I., Fairweather N.F., Charles I.G., Watts C., Hewitt E. and Isaacs N.W. 2000. The Structures of the H(C) Fragment of Tetanus Toxin with Carbohydrate Subunit Complexes Provide Insight into Ganglioside Binding *J Biol Chem* 275:8889-8894.

Encinar J.A., Fernandez A., Ferragut J.A., Gonzalez-Ros J.M., DasGupta B.R., Montal M. and Ferrer-Montiel A. 1998. Structural stabilization of botulinum neurotoxins by tyrosine phosphorylation *FEBS Lett* 429:78-82.

Eniola A.O., Rodgers S.D. and Hammer D.A. 2002. Characterization of biodegradable drug delivery vehicles with the adhesive properties of leukocytes *Biomaterials* 23:2167-2177.

Eswaramoorthy S., Kumaran D. and Swaminathan S. 2002. A Novel Mechanism for Clostridium botulinum Neurotoxin Inhibition *Biochemistry* 41:9795-9802.

Evans P.R. 1999. Some notes on choices in data collection *Acta Crystallogr D* 55:1771-1772.

Facchiano F. and Luini A. 1992. Tetanus toxin potently stimulates tissue transglutaminase: A possible mechanism of neurotoxicity *J Biol Chem* 267:13267-13271.

Fasshauer D., Eliason W.K., Brunger A.T. and Jahn R. 1998. Identification of a Minimal Core of the Synaptic SNARE Complex Sufficient for Reversible Assembly and Disassembly *Biochemistry* 37:10354-10362.

Ferrer-Montiel A., Canaves J.M., DasGupta B.R., Wilson M.C. and Montal M. 1996. Tyrosine Phosphorylation Modulates the Activity of Clostridial Neurotoxins *J Biol Chem* 271:18322-18325.

Flicker P.F., Robinson J.P. and DasGupta B.R. 1999. Is Formation of Visible Channels in Phospholipid Bilayers by Botulinum Neurotoxin Type B Sensitive to Its Disulfide? *J Struct Biol* 128:297-304.

Foran P.G., Mohammed N., Lisk G.O., Nagwaney S., Lawrence G.W., Johnson E., Smith L., Aoki K.R. and Dolly J.O. 2002. Evaluation of the Therapeutic Usefulness of Botulinum Neurotoxin B, C1, E & F Compared to the Long-Lasting Type A *J Biol Chem* 278:1363-1371.

Fu F.N., Lomneth R.B., Cai S. and Singh B.R. 1998. Role of Zinc in the Structure and Toxic Activity of Botulinum Neurotoxin *Biochemistry* 37:5267-5278.

Fujinaga Y., Inoue K., Watanabe S., Yokota K., Hirai Y., Nagamachi E. and Oguma K. 1997. The haemagglutinin of Clostridium botulinum type C progenitor toxin plays an essential role in binding of toxin to the epithelial cells of guniea pig small intestine, leading to the efficient absorption of the toxin *Microbiology* 143:3841-3847.

Gabor F., Stangl M. and Wirth M. 1998. Lectin-mediated bioadhesion binding characteristics of plant lectins on the enterocyte-like cell lines Caco-2, HT-29 and HCT-5 *J Control Release* 55:131-142.

Garman E. 1999. Cool data: quantity AND quality *Acta Crystallogr D* 55:1641-1653.

Georgiou G. and Valax P. 1996. Expression of correctly folded proteins in Escherichia coli *Curr Opin Biotechnol* 7:190-197.

Gerona R.R.L., Larsen E.C., Kowalchyk J.A. and Martin T.F.J. 2000. The C Terminus of SNAP-25 is Essential for Ca2+-dependent Binding of Synaptotagmin to SNARE Complexes *J Biol Chem* 275:6328-6336.

Gewurz H., Zhang X.H. and Lint T.F. 1995. Structure and function of the pentraxins *Curr Opin Immunol* 7:54-64.

Giacovazzo C. 1994. Symmetry in Crystals. In: Giacovazzo C. ed *Fundamentals of Crystallography*, Oxford University Press

Gill S.C. and von Hippel P.H. 1989. Calculation of protein extinction coefficients from amino acid sequence data *Anal Biochem* 182:319-326.

Ginalski K., Venclovas C., Lesyng B. and Fidelis K. 2000. Structure-based sequence alignment for the beta-trefoil subdomain of the clostridial neurotoxin family provides residue level information about the putative ganglioside binding site *FEBS Lett* 482:119-124.

Goldstein I.J. and Hayes C.E. 1976. The lectins: carbohydrate-binding proteins of plants and animals *Adv Carbohydr Chem Biochem* 35:127-340.

Goodnough M.C., Oyler G., Fishman P.S., Johnson E.A., Neale E.A., Keller J.E., Tepp W.H., Clarke M., Hartz S. and Adler M. 2002. Development of a delivery vehicle for intracellular transport of botulinum neurotoxin antagonists *FEBS Lett* 513:163-168.

Green D.W., Ingram V.M. and Perutz M.F. 1954. The structure of haemoglobin IV. Sign determination by the isomorphous replacement method *Proc Roy Sci A* 225:287-307.

Griffiths G., Hoflack B., Simons K., Mellman I. and Kornfeld S. 1988. The mannose 6-phosphate receptor and the biogenesis of lysosomes *Cell* 52:329-341.

Grosse-Kunstleve R.W. and Adams P.D. 2001. Patterson correlation methods: a review of molecular replacement with CNS *Acta Crystallogr D* 57:1390-1396.

Grupp C., Hemprich U., John H., Braun F., Lorf T., Armstrong V.W., Sattler B., Ringe B. and Muller G.A. 2002. Lectin staining for urine cytologic monitoring after kidney transplantation *Nephrol Dial Transplant* 17:1491-1496.

Grupp C., John H., Hemprich U., Singer A., Munzel U. and Muller G.A. 2001. Identification of nucleated cells in urine using lectin staining *Am J Kidney Dis* 37:84-93.

Hallis B., James B.A.F. and Shone C.C. 1996. Development of Novel Assays for Botulinum Type A and B Neurotoxins Based on Their Endopeptidase Activities *J Clin Microbiol* 34:1934-1938.

Hambleton P. 1992. Clostridium botulinum toxins: a general review of involvement in disease, structure, mode of action and preparation for clinical use *J Neurol* 239:16-20.

Hamblin J. and Kent S.P. 1973. Possible role of phytohaemagglutinin in Phaseolus vulgaris L. *Nat New Biol* 245:28-29.

Hamelryck T.W., Loris R., Bouckaert J., Dao-Thi M.H., Strecker G., Imberty A., Fernandez E., Wyns L. and Etzler M.E. 1999. Carbohydrate binding in quarternary structure and a novel hydrophobic binding site in two legume lectin oligomers from Dolichos biflorus *J Mol Biol* 286:1161-1177.

Hanson M.A. and Stevens R.C. 2000. Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0A resolution *Nat Struct Biol* 7:687-692.

Hardman K.D. and Ainsworth C.F. 1972. Structure of concanavalin A at 2.4-A resolution *Biochemistry* 11:4910-4919.

Hayashi T., McMahon H., Yamasaki S., Binz T., Hata Y., Sudhof T.C. and Niemann H. 1994. Synaptic vesicle membrane fusion complex: action of clostridial neurotoxins on assembly *EMBO J* 13:5051-5061.

Heckmann M., Ceballos-Baumann A.O. and Plewig G. 2001. Botulinum Toxin A for Axillary Hyperhidrosis (Excessive Sweating) *N Engl J Med* 344:488-493.

Helliwell J.R. 1997. Overview of Synchrotron Radiation and Macromolecular Crystallography *Methods Enzymol* 276:203-217.

Hendrickson W.A. and Ogata C.M. 1997. Phase Determination from Multiwavelength Anomalous Diffraction Measurements *Methods Enzymol* 276:494-523.

Herreros J., Lalli G., Montecucco C. and Schiavo C. 2000. Tetanus Toxin Fragment C Binds to a Protein Present in Neuronal Cell Lines and Motoneurons *J Neurochem* 74:1941-1950.

Hille-Rehfeld A. 1995. Mannose 6-phosphate receptors in sorting and transport of lysosomal enzymes *Biochim Biophys Acta* 1241:177-194.

Hirsch A.M., Lum M.R. and Downie J.A. 2001. What Makes the Rhizobia-Legume Symbiosis So Special? *Plant Physiol* 127:1484-1492.

Hoch D.H., Romero-Mira M., Ehrlich B.E., Finkelstein A., DasGupta B.R. and Simpson L.L. 1985. Channels formed by botulinum, tetanus and diphtheria toxins in planar lipid bilayers: Relevance to translocation of proteins across membranes *Proc Natl Acad Sci USA* 82:1692-1696.

Hohne-Zell B., Stecher B. and Gratzi M. 1993. Functional characterization of the catalytic site of the tetanus toxin light chain using permeabilized adrenal chromaffin cells *FEBS Lett* 336:175-180.

Holmskov U., Thiel S. and Jensenius J.C. 2003. COLLECTINS AND FICOLINS: Humoral Lectins of the Innate Immune Defense *Annu Rev Immunol* 21:547-578.

Humeau Y., Doussau F., Grant N.J. and Poulain B. 2000. How botulinum and tetanus neurotoxins block neurotransmitter release *Biochimie* 82:427-446.

Iglesias J.L., Lis H. and Sharon N. 1982. Purification and Properties of a D-Galactose/N-Acetyl-D-galactosamine-Specific Lectin from Erythrina cristagalli *Eur J Biochem* 123:247-252.

Ikeda K., Sannoh T., Kawasaki N., Kawasaki T. and Yamashina I. 1987. Serum Lectin with Known Structure Activates Complement through the Classical Pathway *J Biol Chem* 262:7451-7454.

Jacobson S., Carstensen A., Kallenius G. and Svenson S. 1986. Fluorescence-activated cell analysis of P-fimbriae receptor accessibility on uroepithelial cells of patients with renal scarring *Eur J Clin Microbiol* 5:649-654.

Jaffe C.L., Ehrlich-Rogozinski S., Lis H. and Sharon N. 1977. Transition Metal Requirements of Soybean Agglutinin *FEBS Lett* 82:191-196.

Jahn R. and Sudhof T.C. 1999. Membrane Fusion and Exocytosis *Ann Rev Biochem* 68:863-911.

Johnson E.A. 1999. Clostridial Toxins as Therapeutic AgentsL: Benefits of Nature's Most Toxic Proteins *Ann Rev Microbiol* 53:551-575.

Jones T.A. and Kjeldgaard M. 1997. Electron-Density Map Interpretation *Methods Enzymol* 277:173-208.

Jones T.A., Zou J.Y., Cowan S.W. and Kjeldgaard M. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models *Acta Crystallogr A* 47:110-119.

Julius D. and Basbaum A.I. 2001. Molecular mechanisms of nociception *Nature* 413:203-210.

Kadkhodayan S., Knap M.S., Schmidt J.J., Fabes S.E., Rupp B. and Balhorn R. 2000. Cloning, Expression, and One-Step Purification of the Minimal Essential Domain of the Light Chain of Botulinum Neurotoxin Type A *Prot Expr Purif* 19:125-130.

Kagan B.L., Finkelstein A. and Colombini M. 1981. Diphtheria toxin fragment forms large pores in phoshplipids bilayer membranes *Proc Natl Acad Sci USA* 78:4950-4954.

Kalandakanond S. and Coffield J.A. 2001. Cleavage of SNAP-25 by Botulinum Toxin Type A Requires Receptor-Mediated Endocytosis, pH-Dependent Translocation and Zinc *J Pharm Exp Ther* 296:980-986.

Karle J. 1989. Direct methods in protein crystallography *Acta Crystallogr A* 45:765-781.

Ke H. 1997. Overview of Isomorphous Replacement Phasing *Methods Enzymol* 277:208-230.

Kelly S.M. and Price N.C. 1997. The application of circular dichroism to studies of protein folding and unfolding *Biochim Biophys Acta* 1338:161-185.

Kelm S., Brossmer R., Isecke R., Gross H.J., Strenge K. and Schauer R. 1998. Functional groups of sialic acids involved in binding to siglecs (sialoadhesins) deduced from interactions with synthetic analogues *Eur J Biochem* 255:663-672.

Kiyatkin N., Maksymowych A.B. and Simpson L.L. 1997. Induction of an Immune Response by Oral Administration of Recombinant Botulinum Toxin *Infect Immun* 65:4586-4591.

Kleywegt G.J. and Brunger A.T. 1996. Checking your imagination: applications of the free R value *Structure* 4:897-904.

Kleywegt G.J. and Jones T.A. 1996. xdlMAPMAN and xdlDATAMAN - Programs for Reformatting, Analysis and Manipulation of Biomacromolecular Electron-DEnsity Maps and Reflection Data Sets *Acta Crystallogr D* 52:826-828.

Kleywegt G.J. and Jones T.A. 1997. Model Building and Refinement Practice *Methods Enzymol* 277:208-230.

Kleywegt G.J. and Jones T.A. 1998. Databases in protein crystallography *Acta Crystallogr D* 54:1119-1131.

Koriazova L.K. and Montal M. 2003. Translocation of botulinum neurotoxin light chain protease through the heavy chain channel *Nat Struct Biol* 10:13-18.

Kornfeld S. and Mellman I. 1989. The Biogenesis of Lysosomes *Annu Rev Cell Biol* 5:483-525.

Kozaki S., Kamata Y., Wataral S., Nishiki T. and Mochida S. 1998. Ganglioside GT1b as a complementary receptor component for Clostridium botulinum neurotoxins *Microbial Pathogenesis* 25:91-99.

Kraulis P.J. 1991. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures *J Appl Cryst* 24:946-950.

Krieg J., Glasner W., Vicentini A., Doucey M.A., Loffler A., Hess D. and Hofsteenge J. 1997. C-Mannosylation of human RNase 2 is an intracellular process performed by a variety of cultured cells *J Biol Chem* 272:26687-26692.

Kuwabara I. and Liu F.T. 1996. Galectin-3 promotes adhesion of human neutrophils to laminin *J Immunol* 156:3939-3944.

Lacy D.B. and Stevens R.C. 1997. Recombinant expression and purification of the botulinum neurotoxin type A translocation domain *Prot Expr Purif* 11:195-200.

Lacy D.B. and Stevens R.C. 1999. Sequence Homology and Structural Analysis of the Clostridial Neurotoxins *J Mol Biol* 291:1091-1104.

Lacy D.B., Tepp W., Cohen A.C., DasGupta B.R. and Stevens R.C. 1998. Crystal structure of botulinum neurotoxin type A and implications for toxicity *Nat Struct Biol* 5:898-902.

Lang T., Bruns D., Wenzel D., Riedel D., Holroyd P., Thiele C. and Jahn R. 2001. SNAREs are concentrated in cholesterol-dependent clusters that define docking and fusion sites for exocytosis *EMBO J* 20:2202-2213.

LaPenotiere H.F., Clayton M.A. and Middlebrook J.L. 1995. Expression of a large, non-toxic fragment of Botulinum Neurotoxin serotype A and its use as an immunogen *Toxicon* 33:1383-1386.

Laskowski R.A., MacArthur M.W., Moss D.S. and Thornton J.M. 1993. PROCHECK: a program to check the stereochemical quality of protein structures *J Appl Cryst* 26:283-291.

Lasky L.A. 1991. Lectin cell adhesion molecules (LEC-CAMs): a new family of cell adhesion proteins involved with inflammation *J Cell Biochem* 45:139-146.

Lebeda F.J. and Olson M.A. 1994. Secondary Structural Predictions for the Clostridial Neurotoxins *Proteins: Struct Funct Genet* 20:293-300.

Lee Y.C. and Lee R.T. 1995. Carbohydrate-Protein Interactions: Basis of Glycobiology *Acc Chem Res* 28:321-327.

Lehr C.M. 2000. Lectin-mediated drug delivery: The second generation of bioadhesives *J Control Release* 65:19-29.

Lencer W.I., Constable C., Moe S., Rufo P.A., Wolf A., Jobling M.G., Ruston S.P., Madara J.L., Holmes R.K. and Hirst T.R. 1997. Proteolytic Activation of Cholera Toxin and Escherichia coli Labile Toxin by Entry into Host Epithelial Cells *J Biol Chem* 272:15562-15568.

Lesley S.A. 2001. High-throughput proteomics: protein expression and purification in the postgenomic world *Prot Expr Purif* 22:159-164.

Leslie A.G.W. 1999. Integration of macromolecular diffraction data *Acta Crystallogr D* 55:1696-1702.

Li L., Binz T., Niemann H. and Singh B.R. 2000. Probing the Mechanistic Role of Glutamate Residue in the Zinc-Binding Motif of Type A Botulinum Neurotoxin Light Chain *Biochemistry* 39:2399-2405.

Li L. and Singh B.R. 2000a. Role of Zinc Binding in Type A Botulinum Neurotoxin Light Chain's Toxic Structure *Biochemistry* 39:10581-10586.

Li L. and Singh B.R. 2000b. Spectroscopic Analysis of pH-Induced Changes in the Molecular Features of Type A Botulinum Neurotoxin Light Chain *Biochemistry* 39:6466-6474.

Lis H. and Sharon N. 1987. Erythrina lectins *Methods Enzymol* 138:544-551.

Lis H. and Sharon N. 1998. Lectins: Carbohydrate-Specific Proteins That Mediate Cellular Recognition *Chem Rev* 98:637-674.

Lobley A. and Wallace B.A. 2001. DICHROWEB: A website for the analysis of protein secondary structure from circular dichroism spectra *Biophys J* 80:3373a.

Lobley A., Whitmore L. and Wallace B.A. 2002. DICHROWEB: An interactive website for the analysis of protein secondary structure from circular dichroism spectra *Bioinformatics* 18:211-212.

Loris R. 2002. Principles of structures of animal and plant lectins *Biochim Biophys Acta* 1572:198-208.

Loris R., Hamelryck T., Bouckaert J. and Wyns L. 1998. Legume lectin structure *Biochim Biophys Acta* 1383:9-36.

Loris R., Stas P.P.G. and Wyns L. 1994. Conserved Waters in Legume Lectin Crystal Structures *J Biol Chem* 269:26722-26733.

Lotan R., Sharon N. and Mirelman D. 1975. Interaction of wheat-germ agglutinin with bacterial cells and cell-wall polymers *Eur J Biochem* 55:257-262.

Maksymowych A.B. and Simpson L.L. 1998. Binding and transcytosis of botulinum neurotoxin by polarized human colon-carcinoma cells *J Biol Chem* 273

Manavalan P. and Johnson W.C.J. 1987. Variable selection method improves the prediction of protein secondary structure from circular dichroism spectra *Anal Biochem* 167:76-85.

Manoj N. and Suguna K. 2001. Signature of quarternary structure in the sequences of legume lectins *Prot Eng* 14:735-745.

Matteoli M., Verderio C., Rossetto O., Iezzi N., Coco S., Schiavo C. and Montecucco C. 1996. Synaptic vesicle endocytosis mediates the entry of tetanus neurotoxin into hippocampal neurons *Proc Natl Acad Sci USA* 93:13310-13315.

Matthews B.W. 1968. Solvent Content of Protein Crystals *J Mol Biol* 33:491-497.

McCarroll L. and King L.A. 1997. Stable insect cell cultures for recombinant protein production *Curr Opin Biotechnol* 8:590-594.

McPherson A. 2001. A Comparison of Salts for the Crystallisation of Macromolecules *Prot Sci* 10:418-422.

McRee D.E. 1993. *Practical Protein Crystallography*, Academic Press Inc.

Meunier F.A., Lisk G., Sesardic D. and Dolly J.O. 2003. Dynamics of motor nerve terminal remodelling unveiled using SNARE-cleaving botulinum toxins: the extent and duration are dictated by the sites of SNAP-25 truncation *Mol Cell Neurosci* 22:454-466.

Mock W.L. and Stanford D.J. 1996. Arazoformyl dipeptide substrates for thermolysin. Confirmation of a reverse protonation catalytic mechanism *Biochemistry* 35:7369-7377.

Monaco H.L. 1994. Experimental methods in X-ray crystallography. In: Giacovazzo C. ed *Fundamentals of Crystallography*, Oxford University Press

Montecucco C. 1986. How do tetanus and botulinum toxins bind to neuronal membranes? *Trends Biochem Sci* 11:314-317.

Montecucco C., Papini E. and Schiavo G. 1994. Bacterial protein toxins penetrate cells via a four-step mechanism *FEBS Lett* 346:92-98.

Montecucco C. and Schiavo G. 1993. Tetanus and botulinum neurotoxins: a new group of zinc proteases *Trends Biochem Sci* 18:324-327.

Montecucco C. and Schiavo G. 1995. Structure and function of tetanus and botulinum neurotoxins *Q Rev Biophys* 28:423-472.

Montecucco C., Schiavo G., Tugnoli V. and De Grandis D. 1996. Botulinum neurotoxins: mechanism of action and therapeutic applications *Mol Med Today* 2:418-424.

Moreno E., Teneberg S., Adar R., Sharon N., Karlsson K.A. and Angstrom J. 1997. Redefinition of the Carbohydrate Specificity of Erythrina corallodendron Lectin Based on Solid-Phase Binding Assays and Molecular Modelling of Native and Recombinant Forms Obtained by Site-Directed Mutagenesis *Biochemistry* 36:4429-4437.

Moriishi K., Koura M., Abe N., Fujii N., Fujinaga Y. and Inoue K. 1996a. Mosaic structures of neurotoxins produced from Clostridium botulinum types C and D organisms *Biochim Biophys Acta* 1307:123-126.

Moriishi K., Koura M., Fujii N., Fujinaga Y., Inoue K., Syuto B. and Oguma K. 1996b. Molecular Cloning of the Gene Encoding the Mosaic Neurotoxin , Composed of Parts of Botulinum Neurotoxin Types C1 and D, and PCR Detectoin of This Gene from Clostridium botulinum Type C Organisms *Appl Env Microbiol* 62:662-667.

Morris A.L., MacArthur M.W., Hutchinson E.G. and Thornton J.M. 1992. Stereochemical quality of protein structure coordinates *Proteins: Struct Funct Genet* 12:345-364.

Muchmore S.W. 1999. Experiences with CCD detectors on a home X-ray source *Acta Crystallogr D* 55:1669-1671.

Munchau A. and Bhatia K.P. 2000. Uses of botulinum toxin injection in medicine today *Br Med J* 320:161-165.

Munro P., Kojima H., Dupont J.-L., Bossu J.-L., Poulain B. and Boquet P. 2001. High Sensitivity of Mouse Neuronal Cells to Tetanus Toxin Requires a GPI-Anchored Protein *Biochem Biophys Res Commun* 289:623-629.

Naumann M., Jost W.H. and Toyka K.V. 1999. Botulinum Toxin in the Treatment of Neurological Disorders of the Autonomic Nervous System *Arch Neurol* 56:914-916.

Navaza J. 1994. AMoRe: an automated package for molecular replacement *Acta Crystallogr A* 50:157-163.

Nave C. 1999. Matching X-ray source, optics and detectors to protein crystallography requirements *Acta Crystallogr D* 55:1663-1668.

Nicholls A., Sharp K. and Honig B. 1991. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons *Proteins: Struct Funct Genet* 11:281-296.

Niemann H. 1991. Molecular Biology of Clostridial Neurotoxins *Sourcebook of Bacterial Protein Toxins*, pp 303-348

Oblatt-Montal M., Yamazaki M., Nelson R. and Montal M. 1995. Formation of ion channels in lipid bilayers by a peptide with the predicted transmembrane sequence of botulinum neurotoxin A *Prot Sci* 4:1490-1497.

Ohishi I., Sugii S. and Sakaguchi G. 1977. Oral toxicities of Clostridium botulinum in response to molecular size *Infect Immun* 16:107-109.

Olsen L.R., Dessen A., Gupta D., Sabesan S., Sacchettini J.C. and Brewer C.F. 1997. X-ray Crystallographic Studies of Unique Cross-Linked Lattices between Four Isomeric Biantennary Oligosaccharides and Soybean Agglutinin *Biochemistry* 36:15073-15080.

Otwinowski Z. and Minor W. 1997. Processing of X-ray Difraction Data Collected in Oscillation Mode *Methods Enzymol* 276:307-326.

Pannu N.S. and Read R.J. 1996. Improved Structure Refinement Through Maximum Likelihood *Acta Crystallogr A* 52:659-668.

Pellegrini L., O'Connor V. and Betz H. 1994. Fusion complex formation protects synaptobrevin against proteolysis by tetanus toxin light chain *FEBS Lett* 353:319-323.

Pellizzari R., Rossetto O., Lozzi L., Giovedi S., Johnson E., Shone C.C. and Montecucco C. 1996. Structural Determinants of the Specificity for Synaptic Vesicle-associated Membrane Protein/Synaptobrevin of Tetanus and Botulinum Type B and G Neurotoxins *J Biol Chem* 271:20353-20358.

Pellizzari R., Rossetto O., Schiavo G. and Montecucco C. 1999. Tetanus and botulinum neurotoxins: mechanism of action and therapeutic uses *Philos Trans R Soc Lond B Biol Sci* 354:259-268.

Perillo N.L., Pace K.E., Seilhamer J.J. and Baum L.G. 1995. Apoptosis of T cells mediated by galectin-1 *Nature* 378:736-739.

Perillo N.L., Uttenbogaart C.H., Nguyen J.T. and Baum L.G. 1997. Galectin-1, an Endogenous Lectin Produced by Thymic Epithelial Cells, Induces Apoptosis of Human Thymocytes *J Exp Med* 185:1851-1858.

Peumans W.J. and van Damme E.J. 1995. The role of lectins in plant defence *Histochem J* 27:253-271.

Pflugrath J.W. 1997. Diffraction-Data Processing for Electronic Detectors: Theory and Practice *Methods Enzymol* 276:286-306.

Poirier M.A., Xiao W., Macosko J.C., Chan C., Shin Y.-K. and Bennett M.K. 1998. The synaptic SNARE complex is a parallel four-stranded helical bundle *Nat Struct Biol* 5:765-769.

Porta M. 2000. A comparative trial of botulinum toxin type A and methylprednisolone for the treatment of myofascial pain syndrome and pain from chronic muscle spasm *Pain* 85:101-105.

Prabu M.M., Sankaranarayanan R., Puri K.D., Sharma V., Surolia A., Vijayan M. and Suguna K. 1998. Carbohydrate Specificity and Quarternary Association in Basic

Winged Bean Lectin: X-ray Analysis of the Lectin at 2.5A Resolution *J Mol Biol* 276:787-796.

Prabu M.M., Suguna K. and Vijayan M. 1999. Variability in Quarternary Association of Proteins With the Same Tertiary Fold: A Case Study and Rationalization Involving Legume Lectins *Proteins: Struct Funct Genet* 35:58-69.

Provencher S.W. and Glockner J. 1981. Estimation of globular protein secondary structure from circular dichroism *Biochemistry* 20:33-37.

Ratcliffe D., Migliorisi G. and Cramer E. 1992. Translocation of influenza virus by migrating neutrophils *Cell Mol Biol* 38:63-70.

Read R.J. 1986. Improved Fourier coefficients for maps using phases from partial structures with errors *Acta Crystallogr A* 42:140-149.

Reeke G.N.J., Becker J.W. and Edelman G.M. 1975. The covalent and Three-Dimensional Structure of Concanavalin A *J Biol Chem* 259:1525-1547.

Rhodes G. 2000. *Crystallography Made Crystal Clear* Second Edition, Academic Press.

Rice L.M. and Brunger A.T. 1994. Torsion Angle Dynamics: Reduced Variable Conformational Sampling Enhances Crystallographic Structure Refinement *Proteins: Struct Funct Genet* 19:277-290.

Riedel W. and Neeck G. 2001. Nociception, pain, and antinociception:current concepts *Z Rheumatol* 60:404-415.

Rigoni M., Caccin P., Johnson E.A., Montecucco C. and Rossetto O. 2001. Site-directed Mutagenesis Identifies Active-Site Residues of the Light Chain of Botulinum Neurotoxin Type A *Biochem Biophys Res Commun* 288:1231-1237.

Rini J.M. and Lobsanov Y.D. 1999. New animal lectin structures *Curr Opin Struct Biol* 9:578-584.

Robinson R.F. and Nahata M.C. 2003. Management of Botulism *Ann Pharmacother* 37:127-131.

Rossetto O., Caccin P., Rigoni M., Tonello F., Bortoletto N., Stevens R.C. and Montecucco C. 2001. Active-site mutagenesis of tetanus neurotoxin implicates Tyr-375 and Glu-271 in metalloproteolytic activity *Toxicon* 39:1151-1159.

Rossetto O., Schiavo G., Montecucco C., Poulain B., Deloye F., Lozzi L. and Shone C.C. 1994. SNARE motif and neurotoxins *Nature* 372:415-416.

Rossmann M.G. and Blow D.M. 1967. A method of positioning a known molecule in an unknown crystal structure *Acta Crystallogr* 23:544-548.

Rossmann M.G. and van Beek C.G. 1999. Data processing *Acta Crystallogr D* 55:1631-1640.

Rummel A., Bade S., Alves J., Bigalke H. and Binz T. 2003. Two Carbohydrate Binding Sites in the $HC_C$-domain of Tetanus Neurotoxin are Required for Toxicity *J Mol Biol* 326:835-847.

Sagane Y., Watanabe T., Kouguchi H., Sunagawa H., Inoue K., Fujinaga Y., Oguma K. and Ohyama T. 2000. Characterization of Nicking of the Nontoxic-Nonhaemagglutinating Components of Clostridium botulinum Types C and D Progenitor Toxin *J Prot Chem* 19:575-581.

Savkovic S.D., Koutsouris A. and Hecht G. 1996. Attachment of a Noninvasive Enteric Pathogen, Enteropathogenic Escherichia coli, to Cultured Human Intestinal Epithelial Monolayers Induces Transmigration of Neutrophils *Infect Immun* 64:4480-4487.

Schiavo G., Matteoli M. and Montecucco C. 2000. Neurotoxins Affecting Neuroexocytosis *Physiol Rev* 80:717-766.

Schmid M.F., Robinson J.P. and DasGupta B.R. 1993. Direct visualisation of botulinum neurotoxin-induced channels in phospholipid vesicles *Nature* 364:827-830.

Schmidt J.J., Stafford R.G. and Bostian K.A. 1998. Type A botulinum neurotoxin proteolytic activity: development of competitive inhibitors and implications for substrate specificity at the S1' binding subsite *FEBS Lett* 435:61-64.

Shaanan B., Lis H. and Sharon N. 1991. Structure of a Legume Lectin with an Ordered N-linked Carbohydrate in Complex with Lactose *Science* 254:862-866.

Shapiro R.E., Specht C.D., Collins B.E., Woods A.S., Cotter R.J. and Schnaar R.L. 1997. Identification of a Ganglioside Recognition Domain of Tetanus Toxin Using a Novel Ganglioside Photoaffinity Ligand *J Biol Chem* 272:30380-30386.

Sharma V. and Surolia A. 1997. Analyses of Carbohydrate Recognition by Legume Lectins: Size of Combining Site Loops and their Primary Specificity *J Mol Biol* 267:433-445.

Sharon N. and Lis H. 2002. How proteins bind carbohydrates: lessons from legume lectins *J Agric Food Chem* 50:6586-91.

Sheridan R.E. 1998. Gating and Permeability of Ion Channels Produced by Botulinum Toxin Types A and E in PC12 Cell Membranes *Toxicon* 36:703-717.

Shone C.C., Quinn C.P., Wait R., Hallis B., Fooks S.J. and Hambleton P. 1993. Proteolytic cleavage of synthetic fragments of vesicle-associated membrane protein, isoform-2 by botulinum type B neurotoxin *Eur J Biochem* 217:965-971.

Silberstein S., Mathew N., Saper J. and Jenkins S. 2000. Botulinum Toxin Type A as a Migraine Preventative Treatment *Headache* 40:445-450.

Simpson L.L. 1988. Use of pharmacologic antagonists to deduce commonalities of biologic activity among clostridial neurotoxins *J Pharm Exp Ther* 245:867-872.

Simpson L.L., Maksymowych A.B. and Hao S. 2001. The Role of Zinc Binding in the Biological Activity of Botulinum Toxin *J Biol Chem* 276:27034-27041.

Singh B.R. and DasGupta B.R. 1989a. Molecular topography and secondary structure composition of botulinum neurotoxin types A, B and E *Mol Cell Biochem* 86:87-95.

Singh B.R. and DasGupta B.R. 1989b. Structure of heavy and light chain subunits of type A botulinum neurotoxin analyzed by circular dichroism and fluorescence measurements *Mol Cell Biochem* 85:67-73.

Singh B.R., Li B. and Read D. 1995. Botulinum versus tetanus neurotoxins: why is botulinum neurotoxin but not tetanus neurotoxin a food poison? *Toxicon* 33:1541-1547.

Sliz P., Harrison S.C. and Rosenbaum G. 2003. How does Radiation Damage in Protein Crystals Depend on X-Ray Dose? *Structure* 11:13-19.

Smith B.J., Colman P.M., von Itzstein M., Danylec B. and Varghese J.N. 2001. Analysis of inhibitor binding in influenza virus neuraminidase *Prot Sci* 10:689-696.

Spiro R.G. 2000. Glucose Residues as Key Determinants in the Biosynthesis and Quality Control of Glycoproteins with N-Linked Oligosaccharides *J Biol Chem* 275:35657-35660.

Spiro R.G. 2002. Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds *Glycobiology* 12:43R-56R.

Sreerama N., Venyaminov S.Y. and Woody R.W. 1999. Estimation of the number of helical and strand segments in proteins using CD spectroscopy *Prot Sci* 8:370-380.

Sreerama N., Venyaminov S.Y. and Woody R.W. 2000. Estimation of protein secondary structure from CD spectra: Inclusion of denatured proteins with native proteins in the analysis *Anal Biochem* 287:243-251.

Sreerama N. and Woody R.W. 1993. A self-consistent method for the analysis of protein secondary structure from circular dichroism *Anal Biochem* 209:32-44.

Sreerama N. and Woody R.W. 2000. Estimation of protein secondary structure from CD spectra: Comparison of CONTIN, SELCON and CDSSTR methods with an expanded reference set *Anal Biochem* 287:252-260.

Stancombe P.R., Alexander F.C.G., Ling R.J., Matheson M.A., Shone C.C. and Chaddock J.A. 2003. Isolation of the gene and large scale expression and purification of recombinant Erythrina cristagalli lectin *Prot Expr Purif* 30:283-292.

Stevens R.C., Evenson M.L., Tepp W. and DasGupta B.R. 1991. Crystallization and preliminary X-ray analysis of botulinum neurotoxin type A *J Mol Biol* 222:877-880.

Streit W.J., Schulte B.A., Balentine J.D. and Spicer S.S. 1985. Histochemical localization of galactose-containing glycoconjugates in sensory neurons and their process in the central and peripheral nervous system of the rat *J Histochem Cytochem* 33:1042-1052.

Streit W.J., Schulte B.A., Balentine J.D. and Spicer S.S. 1986. Evidence for glycoconjugate in nociceptive primary sensory neurons and its origin from the Golgi complex *Brain Res* 377:1-17.

Sudbery P.E. 1996. The expression of recombinant proteins in yeasts *Curr Opin Biotechnol* 7:517-524.

Sutton J.M., Chow-Worn O., Spaven L., Silman N.J., Hallis B. and Shone C.C. 2001. Tyrosine-1290 of tetanus neurotoxin plays a key role in its binding to gangliosides and functional binding to neurones *FEBS Lett* 493:45-49.

Svensson C., Teneberg S., Nilsson C.L., Kjellberg A., Schwarz F.P., Sharon N. and Krengel U. 2002. High-resolution crystal structures of Erythrina cristagalli lectin in complex with lactose and 2'-alpha-L-fucosyllactose and correlation with thermodynamic binding data *J Mol Biol* 321:69-83.

Swaminathan S. and Eswaramoorthy S. 2000a. Crystallization and preliminary X-ray analysis of Clostridium botulinum neurotoxin type B *Acta Crystallogr D* 56:1024-1026.

Swaminathan S. and Eswaramoorthy S. 2000b. Structural analysis of the catalytic and binding sites of Clostridium botulinum neurotoxin B *Nat Struct Biol* 7:693-699.

Teneberg S., Angstrom J., Jovall P.-A. and Karlsson K.A. 1994. Characterization of Binding of GalBGlcNAC-specific Lectins from Erythrina cristagalli and Erythrina corallodendron to Glycosphingolipids *J Biol Chem* 269:8554-8563.

Thompson J.D., Higgins D.G. and Gibson T.J. 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice *Nucl Acid Res* 22:4673-4680.

Tonello F., Schiavo G. and Montecucco C. 1997. Metal substitution of tetanus neurotoxin *Biochem J* 322:507-510.

Tsuneoka M., Nakayama K., Hatsuzawa K., Komada M., Kitamura N. and Mekada E. 1993. Evidence for Involvement of Furin in Cleavage and Activation of Diphtheria Toxin *J Biol Chem* 266:26461-26465.

Turco S.J. and Robbins P.W. 1979. The Initial Stages of Processing of Protein-bound Oligosaccharides in Vitro *J Biol Chem* 254:4560-4567.

Turton K., Chaddock J.A. and Acharya K.R. 2002. Botulinum and tetanus neurotoxins: structure, function and therapeutic utility *Trends Biochem Sci* 27:552-558.

Umland T.C., Wingert L.M., Swaminathan S., Furey W.F., Schmidt J.J. and Sax M. 1997. Structure of the receptor binding fragment HC of tetanus toxin *Nat Struct Biol* 4:788-792.

Uzzau S. and Fasano A. 2000. Cross-talk between enteric pathogens and the intestine *Cell Microbiol* 2:83-89.

van den Berg T.K., Breve J.J., Damoiseaux J.G., Dopp E.A., Kelm S., Crocker P.R., Dijkstra C.D. and Kraal G. 1992. Sialoadhesin on macrophages: its identification as a lymphocyte adhesion molecule *J Exp Med* 176:647-655.

van Rhijn P., Goldberg R.B. and Hirsch A.M. 1998. Lotus corniculatus Nodulation Specificity is Changed by the Presence of a Soybean Lectin Gene *The Plant Cell* 10:1233-1249.

Van Stokkum I.H.M., Spoelder H.J.W., Bloemendal M., Van Grondelle R. and Groen F.C.A. 1990. Estimation of protein secondary structure and error analysis from CD spectra *Anal Biochem* 191:110-118.

Verderio C., Coco S., Rossetto O., Montecucco C. and Matteoli M. 1999. Internalization and Proteolytic Action of Botulinum Toxins in CNS Neurons and Astrocytes *J Neurochem* 73:372-379.

Vespa G.N.R., Lewis L.A., Kozak K.R., Moran M., Nguyen J.T., Baum L.G. and Miceli M.C. 1999. Galectin-1 Specifically Modulates TCR Signals to Enhance TCR Apoptosis but Inhibit IL-2 Production and Proliferation *J Immunol* 162:799-806.

von Itzstein M., Dyason J.C., Oliver S.W., White H.F., Wu W.Y., Kok G.B. and Pegg M.S. 1996. A Study of the Active Site of Influenza Virus Sialidase: An Approach to the Rational Design of Novel Anti-Influenza Drugs *J Med Chem* 39:388-391.

Walsh M.A., Evans G., Sanishvili R., Dementieva I. and Joachimiak A. 1999. MAD data collection - current trends *Acta Crystallogr D* 55:1726-1732.

Weber P.C. 1997. Overview of Protein Crystallization Methods *Methods Enzymol* 276:13-22.

Weigel P.H. 1994. Galactosyl and N-acetylgalactosaminyl homeostasis: a function for mammalian asialoglycoprotein receptors *Bioessays* 16:519-524.

Weimbs T., Low S.H., Chapin S.J., Mostov K.E., Bucher P. and Hofmann K. 1997. A conserved domain is present in different families of vesicular fusion proteins: A new superfamily *Proc Natl Acad Sci USA* 94:3046-3051.

Weis W.I. and Scheller R.H. 1998. Membrane fusion. SNARE the rod, coil the complex *Nature* 395:328-329.

Wheeler A.H. 1998. Botulinum Toxin A, Adjunctive Therapy for Refractory Headaches Associated with Pericranial Muscle Tension *Headache* 38:468-471.

Wictome M., Rossetto O., Montecucco C. and Shone C.C. 1996. Substrate residues N-terminal to the cleavage site of botulinum type B neurotoxin play a role in determining the specificity of its endopeptidase activity *FEBS Lett* 386:133-136.

Williamson L.C., Bateman K.E., Clifford J.C.M. and Neale E.A. 1999. Neuronal Sensitivity to Tetanus Toxin Requires Gangliosides *J Biol Chem* 274:25173-25180.

Wirth M., Fuchs A., Wolf M., Ertl B. and Gabor F. 1998. Lectin-mediated drug targeting: preparation, binding characteristics, and antiproliferative activity of wheat germ agglutinin conjugated to doxorubicin on Caco-2 cells *Pharm Res* 15:1031-1037.

Yowler B.C., Kensinger R.D. and Schengrund C.-L. 2002. Botulinum Neurotoxin A Activity is Dependent upon the Presence of Specific Gangliosides in Neuroblastoma Cells Expressing Synaptotagmin I *J Biol Chem* 277:32815-32819.

Zanotti G. 1994. Protein Crystallography. In: Giacovazzo C. ed *Fundamentals of Crystallography*, Oxford University Press, pp 535-594 IUCr Texts on Crystallography

Zdanovskaia M.V., Los G. and Zdanovsky A.G. 2000. Recombinant Derivatives of Clostridial Neurotoxins as Delivery Vehicles for Proteins and Small Organic Molecules *J Prot Chem* 19:699-707.

Zeelen J.P. 1999. Interpretation of the Crystallization Drop Results. In: Bergfors T.M. ed *Protein Crystallization: Techniques, Strategies and Tips*, International University Line, pp 131-138

Zhou L., dePaiva A., Liu D., Aoki R. and Dolly J.O. 1995. Expression and purification of the light chain of botulinum neurotoxin A: a single mutation abolishes its cleavage of SNAP-25 and neurotoxicity after reconstruction with the heavy chain *Biochemistry* 34:15175-15181.

# APPENDIX ONE
# *ERYTHRINA CRISTAGALLI* LECTIN

**EXPRESSION AND PURIFICATION OF RECOMBINANT ECL**

**RAMACHANDRAN PLOTS FOR CRYSTAL STRUCTURES OF ECL**

**Flowchart Outlining the Steps Involved in the Recombinant Expression and Purification of ECL**

(as described in Stancombe et al, 2003)

*E.cristagalli* seeds were germinated to produce saplings

↓

Genomic DNA was prepared from leaf material and amplified by PCR

↓

The PCR products were cloned into pCR2.1 vector and transformed into *E.coli* TOP10 cells to create clone pCR-ECL

↓

DNA encoding mature ECL was amplified from pCR-ECL using oligonucleotides containing *Nde*I and *Hind*III cleavage sites

↓

The PCR product was cloned into housing vector pMTL1015 and clone integrity was verified by sequencing

↓

For expression of the MBP-ECL fusion protein, the ECL sequence was amplified from pCR-ECL and cloned into pMal-c2E vector and transformed into *E.coli* lysogenic strain BL21(DE3) cells

↓

Cells were grown in Soytone media (8-litre fermentation scale) for 10-14 hours at 37°C until static growth ($OD_{600}$ = 30)

↓

*E.coli* cell paste was processed in 2-litre equivalent quantities

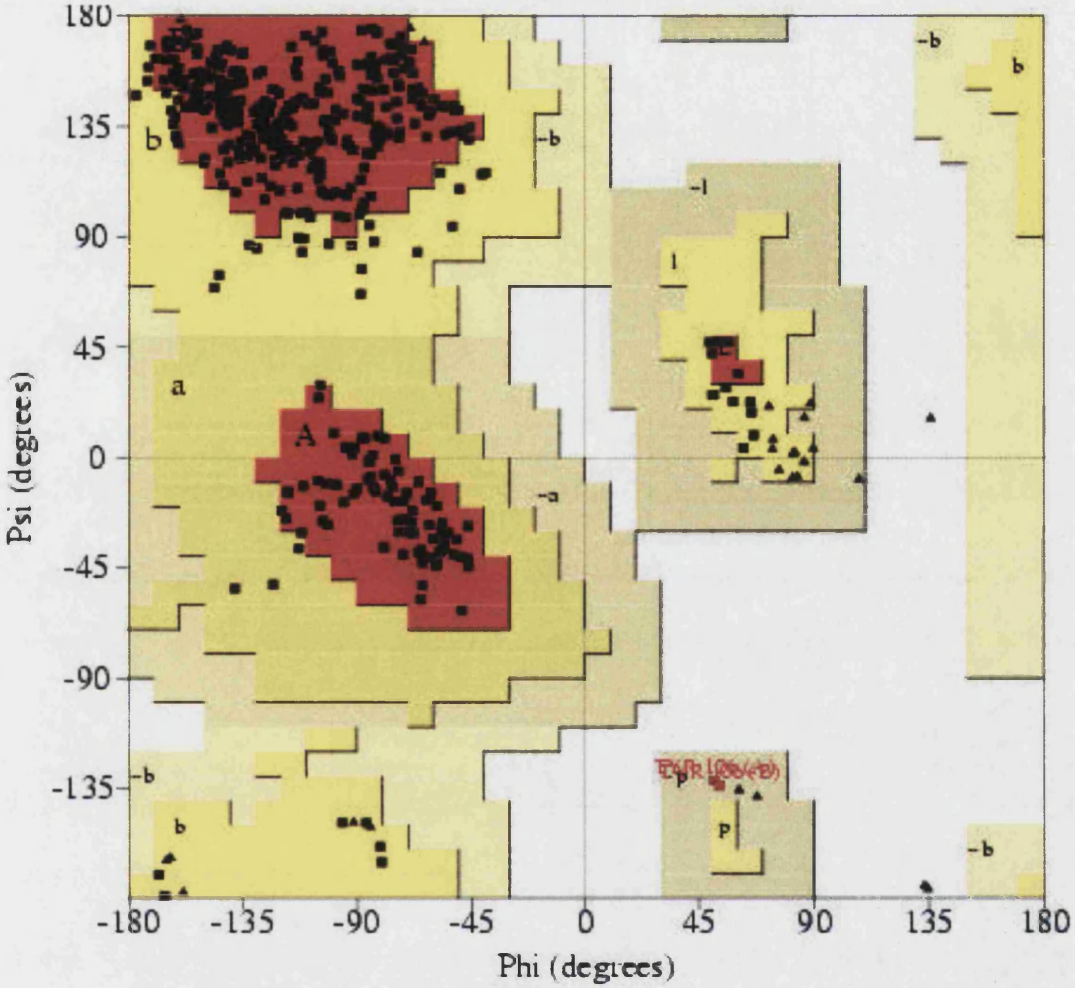
↓

Protein was refolded from inclusion bodies and purified on an immobilised lactose matrix

↓

Pure *rec*ECL was eluted from the matrix with 0.3M lactose

**Ramachandran plot for the Structure of Native ECL in Complex with Lactose at 2.0Å Resolution**

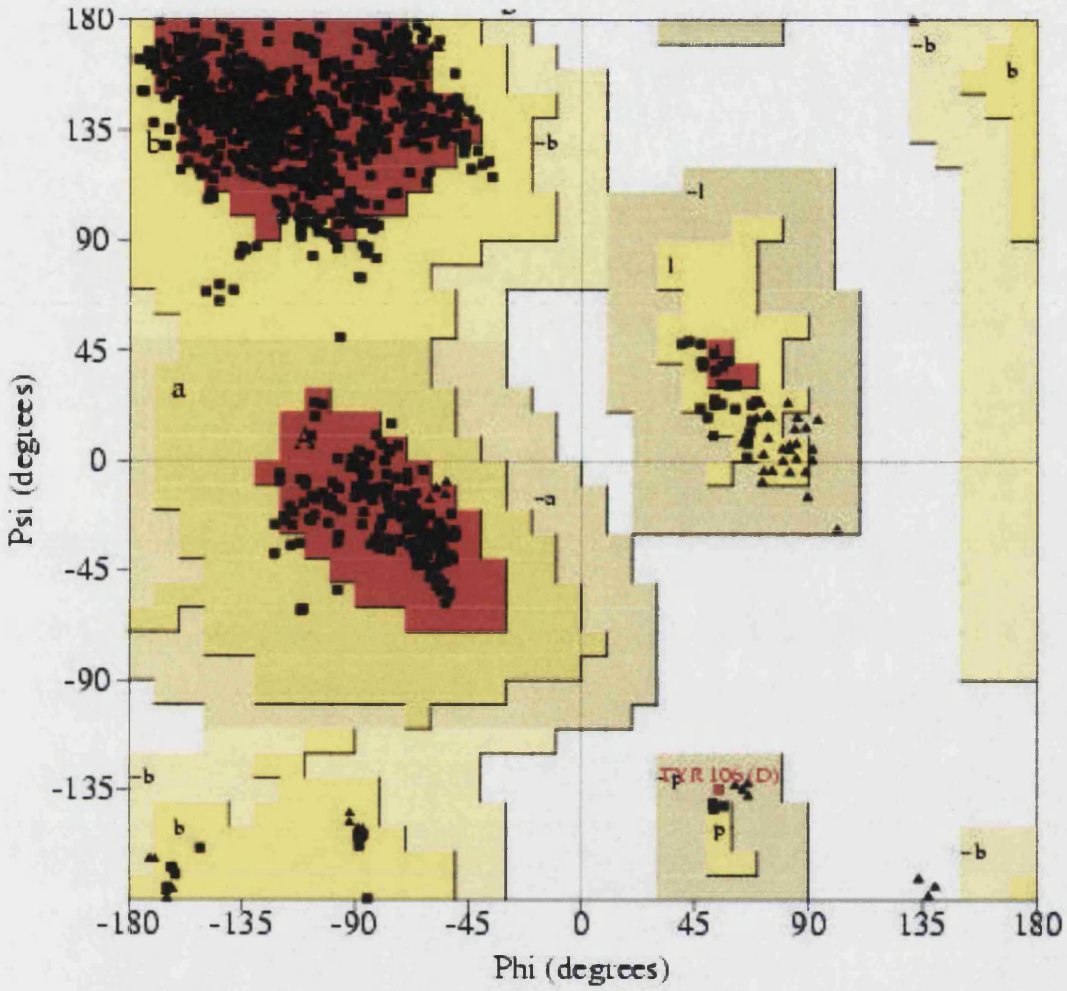



Plot statistics

| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 351 | 86.2% |
| Residues in additional allowed regions [a,b,l,p] | 54 | 13.3% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 2 | 0.5% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 407 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 422 | |
| Number of glycine residues (shown as triangles) | 34 | |
| Number of proline residues | 32 | |
| Total number of residues | 895 | |

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

**Ramachandran Plot for the Structure of Recombinant ECL at 2.13Å Resolution**

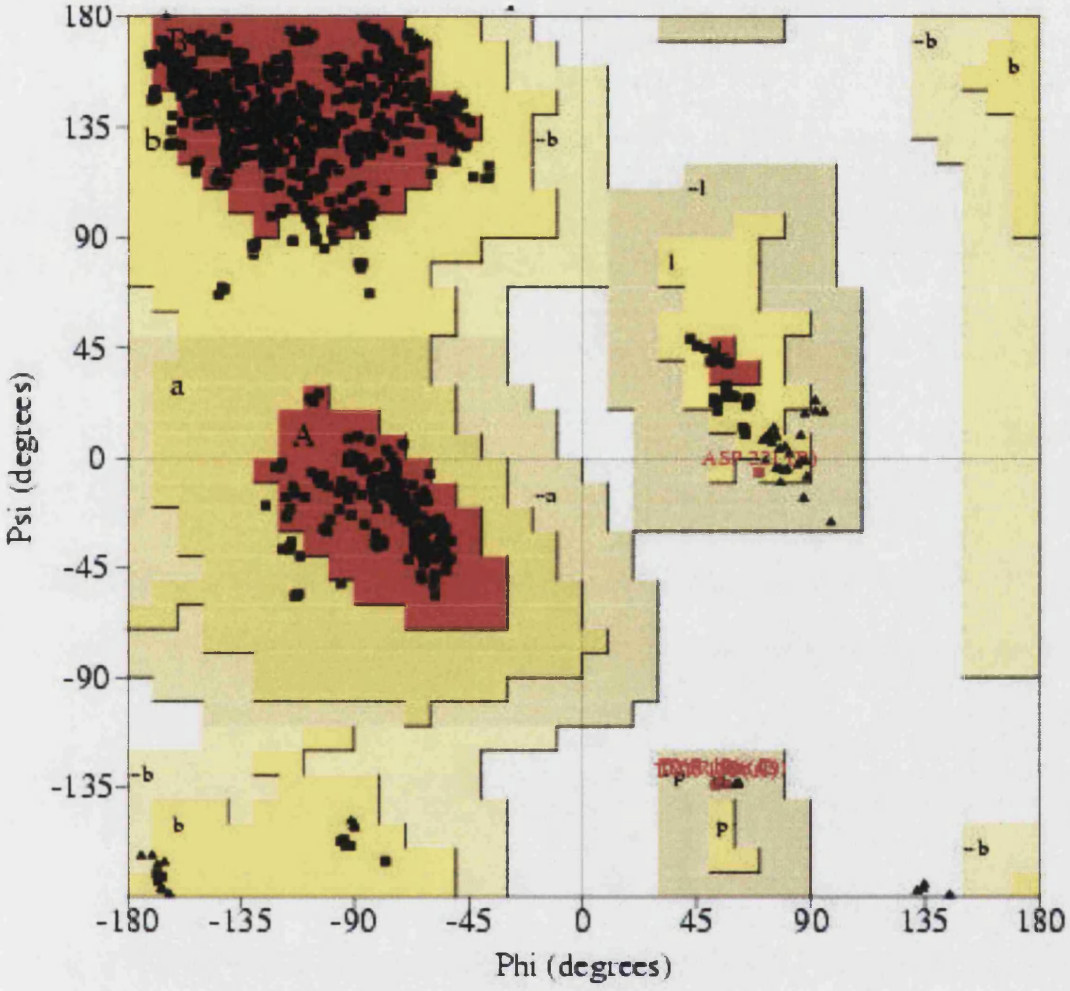



Plot statistics

| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 706 | 86.3% |
| Residues in additional allowed regions [a,b,l,p] | 111 | 13.6% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 1 | 0.1% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 818 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 19 | |
| Number of glycine residues (shown as triangles) | 68 | |
| Number of proline residues | 64 | |
| Total number of residues | 969 | |

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

**Ramachandran Plot for the Structure of Recombinant ECL in Complex with Lactose (Space Group $P2_1$) at 1.7Å Resolution**

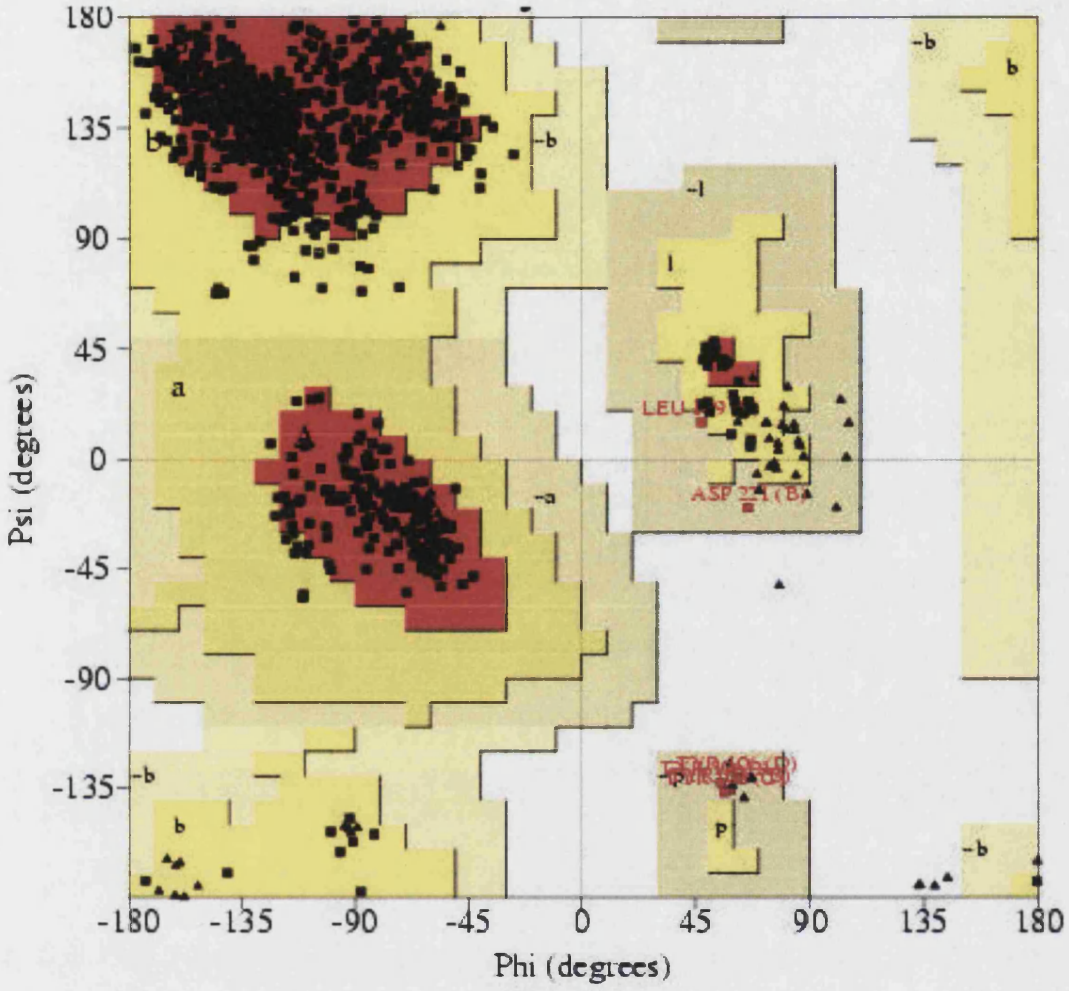



Plot statistics

| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 721 | 88.1% |
| Residues in additional allowed regions [a,b,l,p] | 92 | 11.2% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 5 | 0.6% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 818 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 20 | |
| Number of glycine residues (shown as triangles) | 68 | |
| Number of proline residues | 64 | |
| Total number of residues | 970 | |

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

**Ramachandran Plot for the Structure of Recombinant ECL in Complex with Lactose (Space Group $P2_1$) at 1.68Å Resolution**

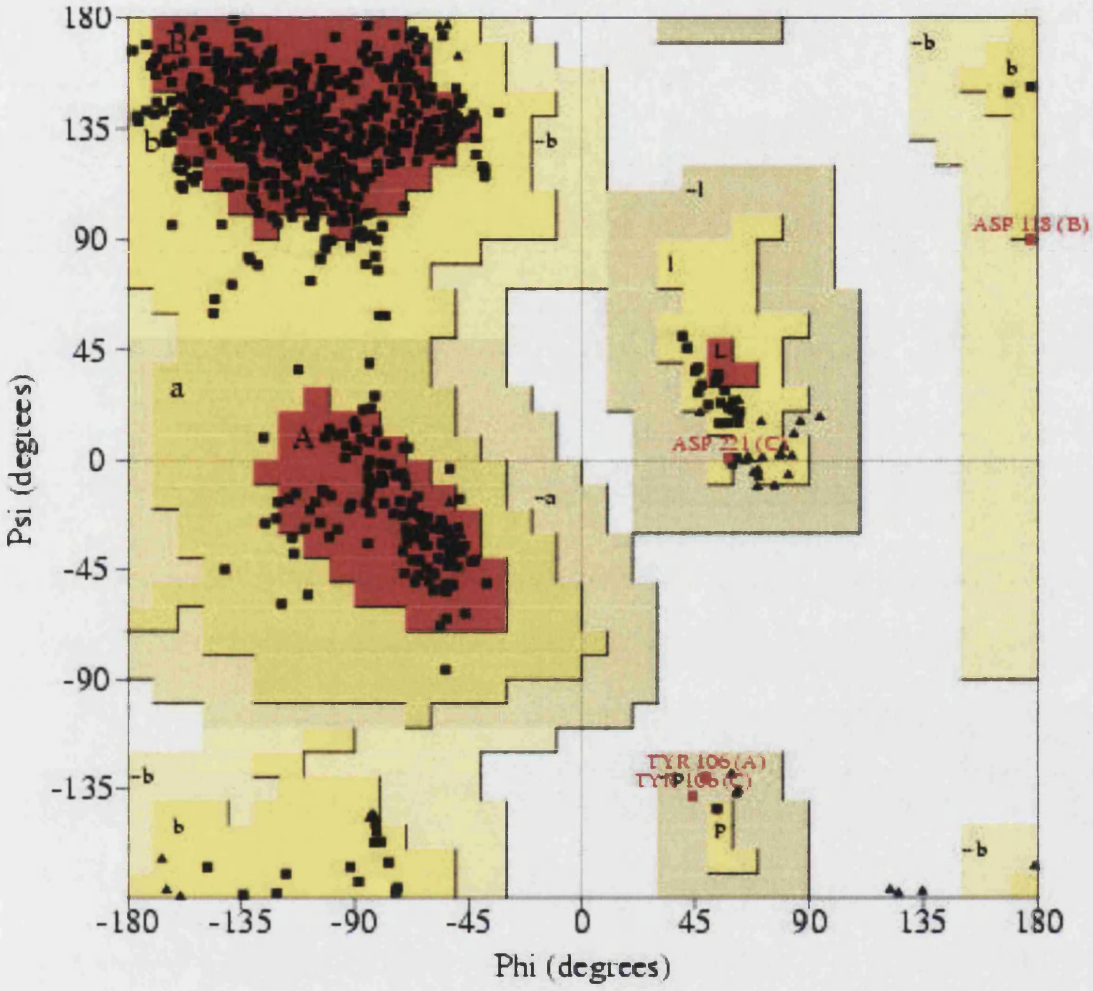



Plot statistics

| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 707 | 87.0% |
| Residues in additional allowed regions [a,b,l,p] | 100 | 12.3% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 6 | 0.7% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 813 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 19 | |
| Number of glycine residues (shown as triangles) | 68 | |
| Number of proline residues | 64 | |
| Total number of residues | 964 | |

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

**Ramachandran Plot for the Structure of Recombinant ECL in Complex with Lactose (Space Group $P3_212$) at 2.3Å Resolution**

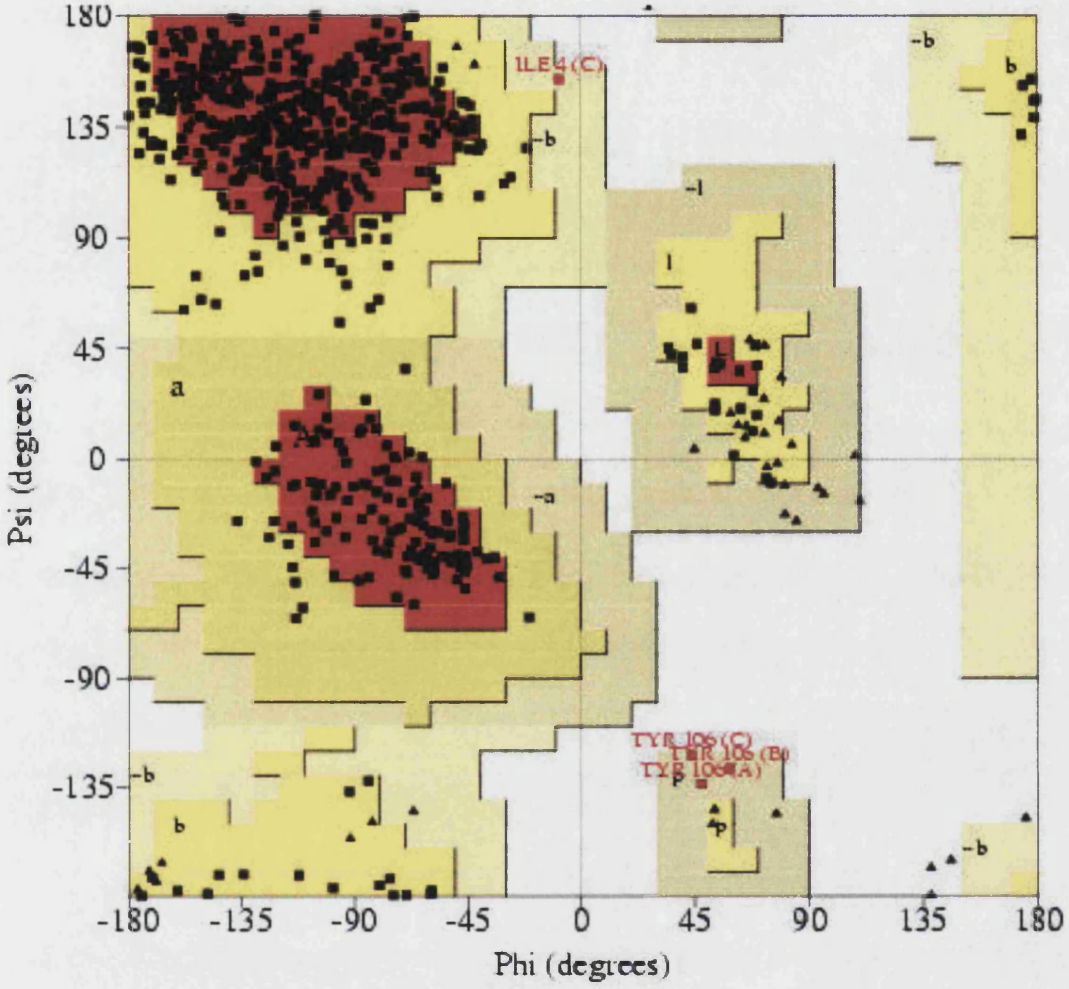



Plot statistics

| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 493 | 81.1% |
| Residues in additional allowed regions [a,b,l,p] | 111 | 18.3% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 4 | 0.7% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 608 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 427 | |
| Number of glycine residues (shown as triangles) | 51 | |
| Number of proline residues | 48 | |
| Total number of residues | 1134 | |

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

**Ramachandran Plot for the Structure of Recombinant ECL in Complex with Lactose (Space Group C2) at 2.45Å Resolution**




Plot statistics

| | | |
|---|---|---|
| Residues in most favoured regions [A,B,L] | 484 | 79.7% |
| Residues in additional allowed regions [a,b,l,p] | 119 | 19.6% |
| Residues in generously allowed regions [~a,~b,~l,~p] | 4 | 0.7% |
| Residues in disallowed regions | 0 | 0.0% |
| Number of non-glycine and non-proline residues | 607 | 100.0% |
| Number of end-residues (excl. Gly and Pro) | 232 | |
| Number of glycine residues (shown as triangles) | 51 | |
| Number of proline residues | 48 | |
| Total number of residues | 938 | |

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

# APPENDIX TWO
# COMMERCIAL CRYSTALLIZATION SCREENS

**STRUCTURE SCREENS**

**PEG/ION MATRIX**

**JBSCREENS**

**CRYSTAL SCREEN**

## Structure Screen 1 - Catalogue Number MD1-01

| | | | |
|---|---|---|---|
| 1 | 0.02M Calcium chloride dihydrate | 0.1M Na Acetate trihydrate pH 4.6 | 30% v/v 2-methyl-2,4 pentanediol |
| 2 | 0.2M Ammonium acetate | 0.1M Na Acetate trihydrate pH 4.6 | 30% w/v PEG 4000 |
| 3 | 0.2M Ammonium sulphate | 0.1M Na acetate trihydrate pH 4.6 | 25% w/v PEG 4000 |
| 4 | | 0.1M Na acetate trihydrate pH 4.6 | 2.0M Sodium formate |
| 5 | | 0.1M Na acetate trihydrate pH 4.6 | 2.0M Ammonium sulphate |
| 6 | | 0.1M Na acetate trihydrate pH 4.6 | 8% w/v PEG 4000 |
| 7 | 0.2M Ammonium acetate | 0.1M tri-sodium citrate dihydrate pH 5.6 | 30% w/v PEG 4000 |
| 8 | 0.2M Ammonium acetate | 0.1M tri-sodium citrate dihydrate pH 5.6 | 30% v/v 2-methyl-2,4-pentanediol |
| 9 | | 0.1M tri-Sodium citrate dihydrate pH 5.6 | 20% w/v 2-propanol 20% w/v PEG 4000 |
| 10 | | 0.1M Na Citrate pH 5.6 | 1.0M Ammonium dihydrogen phosphate |
| 11 | 0.2M Calcium chloride dihydrate | 0.1M Na acetate trihydrate pH 4.6 | 20% v/v 2-propanol |
| 12 | | 0.1M Na Cacodylate pH 6.5 | 1.4M Na acetate trihydrate |
| 13 | 0.2M tri-sodium citrate dihydrate | 0.1M Na Cacodylate pH 6.5 | 30% v/v 2-propanol |
| 14 | 0.2M Ammonium sulphate | 0.1M Na Cacodylate pH 6.5 | 30% w/v PEG 8000 |
| 15 | 0.2M Magnesium acetate tetrahydrate | 0.1M Na Cacodylate pH 6.5 | 20% PEG 8000 |
| 16 | 0.2M Magnesium acetate tetrahydrate | 0.1M Na Cacodylate pH 6.5 | 30% v/v 2-methyl-2,4-pentanediol |
| 17 | | 0.1M Imidazole pH 6.5 | 1.0M Sodium acetate trihydrate |
| 18 | 0.2M Sodium acetate trihydrate | 0.1M Na Cacodylate pH 6.5 | 30% w/v PEG 8000 |
| 19 | 0.2M Zinc acetate dihydrate | 0.1M Na Cacodylate pH 6.5 | 18% w/v PEG 8000 |
| 20 | 0.2M Calcium acetate hydrate | 0.1M Na Cacodylate pH 6.5 | 18% w/v PEG 8000 |
| 21 | 0.2M tri-sodium citrate dihydrate | 0.1M Na Hepes pH 7.5 | 30% v/v 2-methyl-2,4-pentanediol |
| 22 | 0.2M Magnesium chloride hexahydrate | 0.1M Na Hepes pH 7.5 | 30% v/v 2-propanol |
| 23 | 0.2M Calcium chloride dihydrate | 0.1M Na Hepes pH 7.5 | 28% v/v PEG 400 |
| 24 | 0.2M Magnesium chloride hexahydrate | 0.1M Na Hepes pH 7.5 | 30% v/v PEG 400 |
| 25 | 0.2M tri-sodium citrate dihydrate | 0.1M Na Hepes pH 7.5 | 20% v/v 2-propanol |
| 26 | | 0.1M Na Hepes pH 7.5 | 0.8M K, Na tartrate tetrahydrate |
| 27 | | 0.1M Na Hepes pH 7.5 | 1.5M Lithium sulphate monohydrate |
| 28 | | 0.1M Na Hepes pH 7.5 | 0.8M Na dihydrogen phosphate 0.8M K dihydrogen phosphate monohydrate |
| 29 | | 0.1M Na Hepes pH 7.5 | 1.4M tri-Sodium citrate dihydrate |
| 30 | | 0.1M Na Hepes pH 7.5 | 2% v/v PEG 400 2.0M Amm sulphate |
| 31 | | 0.1M Na Hepes pH 7.5 | 10% v/v 2-propanol |

| | | | |
|---|---|---|---|
| | | | 20% w/v PEG 4000 |
| **32** | | 0.1M Tris HCl pH 8.5 | 2.0M Ammoniium sulphate |
| **33** | 0.2M Magnesium chloride hexahydrate | 0.1M Tris HCl pH 8.5 | 30% w/v PEG 4000 |
| **34** | 0.2M tri-sodum citrate dihydrate | 0.1M Tris HCl pH 8.5 | 30% v/v PEG 400 |
| **35** | 0.2M Lithium sulphate monohydrate | 0.1M Tris HCl pH 8.5 | 30% w/v PEG 4000 |
| **36** | 0.2M Ammonium acetate | 0.1M Tris HCl pH 8.5 | 30% v/v 2-propanol |
| **37** | 0.2M Sodium acetate trihydrate | 0.1M Tris HCl pH 8.5 | 30% w/v PEG 4000 |
| **38** | | 0.1M Tris HCl pH 8.5 | 8% w/v PEG 8000 |
| **39** | | 0.1M Tris HCl pH 8.5 | 2.0M Ammonium dihydrogen phosphate |
| **40** | | | 0.4M K, Na Tartrate tetrahydrate |
| **41** | | | 0.4M Ammonium dihydrogen phosphate |
| **42** | 0.2M Ammonium sulphate | | 30% w/v PEG 8000 |
| **43** | 0.2M Ammonium sulphate | | 30% w/v PEG 4000 |
| **44** | | | 2.0M Ammonium sulphate |
| **45** | | | 4.0M Sodium formate |
| **46** | 0.05M Potassium dihydrogen phosphate | | 20% w/v PEG 8000 |
| **47** | | | 30% w/v PEG 1500 |
| **48** | | | 0.2M Magnesium formate |
| **49** | 1.0M Lithium sulphate monohydrate | | 2% w/v PEG 8000 |
| **50** | 0.5M Lithium sulphate monohydrate | | 15% w/v PEG 8000 |

## Structure Screen 2 - Catalogue Number MD1-02

| | | | |
|---|---|---|---|
| 1 | 0.1M Sodium chloride | 0.1M Bicine pH 9.0 | 30% w/v PEG monomethylether 550 |
| 2 | | 0.1M Bicine pH 9.0 | 2.0M Magnesium chloride hexahydrate |
| 3 | 2% w/v Dioxane | 0.1M Bicine pH 9.0 | 10% w/v PEG 20,000 |
| 4 | 0.2M Magnesium chloride hexahydrate | 0.1M Tris pH 8.5 | 3.4M 1,6 Hexanediol |
| 5 | | 0.1M Tris pH 8.5 | 25% v/v tert-Butanol |
| 6 | 0.01M Nickel chloride hexahydrate | 0.1M Tris pH 8.5 | 1.0M Lithium sulphate |
| 7 | 1.5M Ammonium sulphate | 0.1M Tris pH 8.5 | 12% v/v Glycerol |
| 8 | 0.2M Ammonium phosphate monobasic | 0.1M Tris pH 8.5 | 50% v/v MPD |
| 9 | | 0.1M Tris pH 8.5 | 20% v/v Ethanol |
| 10 | 0.01M Nickel chloride hexahydrate | 0.1M Tris pH 8.5 | 20% w/v PEG monomethylether 2000 |
| 11 | 0.5M Ammonium sulphate | 0.1M Hepes pH 7.5 | 30% v/v MPD |
| 12 | | 0.1M Hepes pH 7.5 | 10% w/v PEG 6000<br>5% v/v MPD |
| 13 | | 0.1M Hepes pH 7.5 | 20% v/v Jeffamine M-600 |
| 14 | 0.1M Sodium chloride | 0.1M Hepes pH 7.5 | 1.6M Ammonium sulphate |
| 15 | | 0.1M Hepes pH 7.5 | 2.0M Ammonium formate |
| 16 | 0.05M Cadmium sulphate octahydrate | 0.1M Hepes pH 7.5 | 1.0M Sodium acetate |
| 17 | | 0.1M Hepes pH 7.5 | 70% v/v MPD |
| 18 | | 0.1M Hepes pH 7.5 | 4.3M Sodium chloride |
| 19 | | 0.1M Hepes pH 7.5 | 10% w/v PEG 8000<br>8% v/v Ethylene glycol |
| 20 | | 0.1M Mes pH 6.5 | 1.6M Magnesium sulphate heptahydrate |
| 21 | 0.1M Na phosphate monobasic | 0.1M Mes pH 6.5 | 2.0M Sodium Chloride<br>0.1M K phosphate monobasic |
| 22 | | 0.1M Mes pH 6.5 | 12% w/v PEG 20,000 |
| 23 | 1.6M Ammonium sulphate | 0.1M Mes pH 6.5 | 10% v/v Dioxane |
| 24 | 0.05M Cesium chloride | 0.1M Mes pH 6.5 | 30% v/v Jeffamine M-600 |
| 25 | 0.01M Cobalt chloride hexahydrate | 0.1M Mes pH 6.5 | 1.8M Ammonium sulphate |
| 26 | 0.2M Ammonium sulphate | 0.1M Mes pH 6.5 | 30% w/v PEG monomethylether 5000 |
| 27 | 0.01M Zinc sulphate heptahydrate | 0.1M Mes pH 6.5 | 25% v/v PEG monomethylether 550 |
| 28 | | 0.1M Hepes pH 7.5 | 20% w/v PEG 10,000 |
| 29 | 0.2M K/Na Tartrate | 0.1M Sodium citrate pH 5.6 | 2.0M Ammonium sulphate |
| 30 | 0.5M Ammonium sulphate | 0.1M Sodium citrate pH 5.6 | 1.0M Lithium sulphate |
| 31 | 0.5M Sodium chloride | 0.1M Sodium citrate pH 5.6 | 4% v/v polyethyleneimine |
| 32 | | 0.1M Sodium citrate pH 5.6 | 35% v/v tert-butanol |
| 33 | 0.01M Ferric chloride hexahydrate | 0.1M Sodium citrate pH 5.6 | 10% v/v Jeffamine M-600 |
| 34 | 0.01M Manganese chloride tetrahydrate | 0.1M Sodium citrate pH 5.6 | 2.5M 1,6 Hexanediol |
| 35 | | 0.1M Sodium acetate pH 4.6 | 2.0M Sodium chloride |

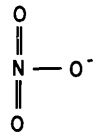
| | | | |
|---|---|---|---|
| **36** | 0.2M Sodium Chloride | 0.1M Sodium acetate pH 4.6 | 30% v/v MPD |
| **37** | 0.01M Cobalt Chloride hexahydrate | 0.1M Sodium acetate pH 4.6 | 1.0M 1,6 Hexanediol |
| **38** | 0.1M Cadmium chloride | 0.1M Sodium acetate pH 4.6 | 30% v/v PEG 400 |
| **39** | 0.2M Ammonium sulphate | 0.1M Sodium acetate pH 4.6 | 30% w/v PEG monomethylether 2000 |
| **40** | 2.0M Sodium Chloride | | 10% w/v PEG 6000 |
| **41** | 0.01M Cetyl trimethyl ammoniumbromide | | 0.5M Sodium chloride<br>0.1M Magnesium chloride hexahydrate |
| **42** | | | 25% v/v Ethylene glycol |
| **43** | | | 35% v/v Dioxane |
| **44** | 2.0M Ammonium Sulphate | | 5% v/v Isopropanol |
| **45** | | | 1.0M Imidazole pH 7.0 |
| **46** | | | 10% w/v PEG 1000<br>10% w/v PEG 8000 |
| **47** | 1.5M Sodium Chloride | | 10% v/v Ethanol |
| **48** | | | 1.6M Sodium citrate pH 6.5 |
| **49** | 15% w/v Polyvinylpyrolidone | | |
| **50** | 2.0M Urea | | |

## PEG/Ion Screen

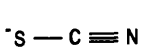
## Reagent Formulation

| Tube Number | Salt | Tube Number | Polyethylene Glycol 3350 | Tube Number | pH |
|---|---|---|---|---|---|
| 1. | 0.2 M Sodium Fluoride | 1. | 20% w/v Polyethylene Glycol 3350 | 1. | 7.1 |
| 2. | 0.2 M Potassium Fluoride | 2. | 20% w/v Polyethylene Glycol 3350 | 2. | 7.2 |
| 3. | 0.2 M Ammonium Fluoride | 3. | 20% w/v Polyethylene Glycol 3350 | 3. | 6.2 |
| 4. | 0.2 M Lithium Chloride anhydrous | 4. | 20% w/v Polyethylene Glycol 3350 | 4. | 6.7 |
| 5. | 0.2 M Magnesium Chloride hexahydrate | 5. | 20% w/v Polyethylene Glycol 3350 | 5. | 5.8 |
| 6. | 0.2 M Sodium Chloride | 6. | 20% w/v Polyethylene Glycol 3350 | 6. | 6.9 |
| 7. | 0.2 M Calicium Chloride dihydrate | 7. | 20% w/v Polyethylene Glycol 3350 | 7. | 5.1 |
| 8. | 0.2 M Potassium Chloride | 8. | 20% w/v Polyethylene Glycol 3350 | 8. | 6.9 |
| 9. | 0.2 M Ammonium Chloride | 9. | 20% w/v Polyethylene Glycol 3350 | 9. | 6.3 |
| 10. | 0.2 M Sodium Iodide | 10. | 20% w/v Polyethylene Glycol 3350 | 10. | 6.9 |
| 11. | 0.2 M Potassium Iodide | 11. | 20% w/v Polyethylene Glycol 3350 | 11. | 6.8 |
| 12. | 0.2 M Ammonium Iodide | 12. | 20% w/v Polyethylene Glycol 3350 | 12. | 6.2 |
| 13. | 0.2 M Sodium Thiocyanate | 13. | 20% w/v Polyethylene Glycol 3350 | 13. | 6.9 |
| 14. | 0.2 M Potassium Thiocyanate | 14. | 20% w/v Polyethylene Glycol 3350 | 14. | 7.0 |
| 15. | 0.2 M Lithium Nitrate | 15. | 20% w/v Polyethylene Glycol 3350 | 15. | 7.1 |
| 16. | 0.2 M Magnesium Nitrate hexahydrate | 16. | 20% w/v Polyethylene Glycol 3350 | 16. | 5.8 |
| 17. | 0.2 M Sodium Nitrate | 17. | 20% w/v Polyethylene Glycol 3350 | 17. | 6.8 |
| 18. | 0.2 M Potassium Nitrate | 18. | 20% w/v Polyethylene Glycol 3350 | 18. | 6.9 |
| 19. | 0.2 M Ammonium Nitrate | 19. | 20% w/v Polyethylene Glycol 3350 | 19. | 6.3 |
| 20. | 0.2 M Magnesium Formate | 20. | 20% w/v Polyethylene Glycol 3350 | 20. | 5.9 |
| 21. | 0.2 M Sodium Formate | 21. | 20% w/v Polyethylene Glycol 3350 | 21. | 7.2 |
| 22. | 0.2 M Potassium Formate | 22. | 20% w/v Polyethylene Glycol 3350 | 22. | 7.3 |
| 23. | 0.2 M Ammonium Formate | 23. | 20% w/v Polyethylene Glycol 3350 | 23. | 6.6 |
| 24. | 0.2 M Lithium Acetate dihydrate | 24. | 20% w/v Polyethylene Glycol 3350 | 24. | 7.8 |
| 25. | 0.2 M Magnesium Acetate tetrahydrate | 25. | 20% w/v Polyethylene Glycol 3350 | 25. | 7.7 |
| 26. | 0.2 M Zinc Acetate dihydrate | 26. | 20% w/v Polyethylene Glycol 3350 | 26. | 6.3 |
| 27. | 0.2 M Sodium Acetate trihydrate | 27. | 20% w/v Polyethylene Glycol 3350 | 27. | 7.9 |
| 28. | 0.2 M Calcium Acetate hydrate | 28. | 20% w/v Polyethylene Glycol 3350 | 28. | 7.3 |
| 29. | 0.2 M Potassium Acetate | 29. | 20% w/v Polyethylene Glycol 3350 | 29. | 7.8 |
| 30. | 0.2 M Ammonium Acetate | 30. | 20% w/v Polyethylene Glycol 3350 | 30. | 7.1 |
| 31. | 0.2 M Lithium Sulfate monohydrate | 31. | 20% w/v Polyethylene Glycol 3350 | 31. | 6.4 |
| 32. | 0.2 M Magnesium Sulfate heptahydrate | 32. | 20% w/v Polyethylene Glycol 3350 | 32. | 5.9 |
| 33. | 0.2 M Sodium Sulfate decahydrate | 33. | 20% w/v Polyethylene Glycol 3350 | 33. | 6.6 |
| 34. | 0.2 M Potassium Sulfate | 34. | 20% w/v Polyethylene Glycol 3350 | 34. | 6.7 |
| 35. | 0.2 M Ammonium Sulfate | 35. | 20% w/v Polyethylene Glycol 3350 | 35. | 6.0 |
| 36. | 0.2 M di-Sodium Tartrate dihydrate | 36. | 20% w/v Polyethylene Glycol 3350 | 36. | 7.2 |
| 37. | 0.2 M Potassium Sodium Tartrate tetrahydrate | 37. | 20% w/v Polyethylene Glycol 3350 | 37. | 7.2 |
| 38. | 0.2 M di-Ammonium Tartrate | 38. | 20% w/v Polyethylene Glycol 3350 | 38. | 6.6 |
| 39. | 0.2 M Sodium dihydrogen Phosphate monohydrate | 39. | 20% w/v Polyethylene Glycol 3350 | 39. | 4.5 |
| 40. | 0.2 M di-Sodium hydrogen Phosphate dihydrate | 40. | 20% w/v Polyethylene Glycol 3350 | 40. | 9.1 |
| 41. | 0.2 M Potassium dihydrogen Phosphate | 41. | 20% w/v Polyethylene Glycol 3350 | 41. | 4.7 |
| 42. | 0.2 M di-Potassium hydrogen Phosphate | 42. | 20% w/v Polyethylene Glycol 3350 | 42. | 9.2 |
| 43. | 0.2 M Ammonium dihydrogen Phosphate | 43. | 20% w/v Polyethylene Glycol 3350 | 43. | 4.6 |
| 44. | 0.2 M di-Ammonium hydrogen Phosphate | 44. | 20% w/v Polyethylene Glycol 3350 | 44. | 7.9 |
| 45. | 0.2 M tri-Lithium Citrate tetrahydrate | 45. | 20% w/v Polyethylene Glycol 3350 | 45. | 8.1 |
| 46. | 0.2 M tri-Sodium Citrate dihydrate | 46. | 20% w/v Polyethylene Glycol 3350 | 46. | 8.2 |
| 47. | 0.2 M tri-Potassium Citrate monohydrate | 47. | 20% w/v Polyethylene Glycol 3350 | 47. | 8.3 |
| 48. | 0.2 M di-Ammonium hydrogen Citrate | 48. | 20% w/v Polyethylene Glycol 3350 | 48. | 5.0 |

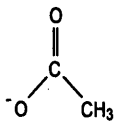

***Fluoride*** ***Chloride*** ***Iodide***



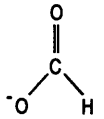
***Nitrate***



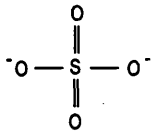
***Thiocyanate***



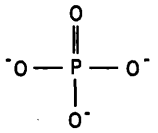
***Acetate***



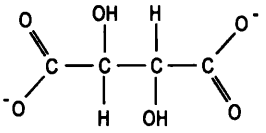
***Formate***



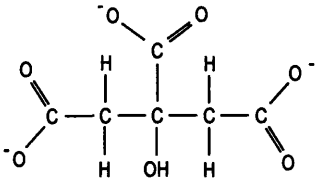
***Sulfate***



***Phosphate***



***Tartrate***



***Citrate***

*PEG/Ion Screen contains forty-eight unique reagents. To determine the formulation of each reagent, simply read across the page.*


34 Journey
Aliso Viejo, CA 92656-3317 U.S.A.
Tel: (949) 425-1321 • Fax: (949) 425-1611
e-mail: tech@hrmail.com
Website: www.hamptonresearch.com


HAMPTON RESEARCH

*Solutions for Crystal Growth*

# JBScreen 7
## (MPD based)

## *Formulations*
**Cat.-No.: CS-107**

| | |
|---|---|
| **A1** | 10% MPD, 0.1 M Na HEPES pH 7.5, 0.1 M Na Citrate |
| **A2** | 12% MPD, 0.1 M TRIS HCl pH 8.5, 0.05 M Mg Chloride |
| **A3** | 15% MPD, 0.1 M Na Acetate pH 4.6, 0.02 M Ca Chloride |
| **A4** | 15% MPD, 5% PEG 4000, 0.1 M Imidazole HCl pH 8.0 |
| **A5** | 15% MPD, 0.1 M Na Citrate pH 5.6, 0.2 M Ammonium Acetate |
| **A6** | 15% MPD, 0.1 M Na MES pH 6.5, 0.2 M Mg Acetate |
| **B1** | 15% MPD, 0.1 M Na HEPES pH 7.5, 0.2 M Na Citrate |
| **B2** | 20% MPD, 0.1 M Na HEPES pH 7.5, 0.1 M Na Citrate |
| **B3** | 20% MPD, 0.1 M Imidazole HCl pH 8.0 |
| **B4** | 20% MPD, 4% Glycerol, 0.2 M Na Chloride |
| **B5** | 30% MPD, 0.1 M Na Acetate pH 4.6, 0.02 M Ca Chloride |
| **B6** | 30% MPD, 0.1 M Na Citrate pH 5.6, 0.2 M Ammonium Acetate |
| **C1** | 30% MPD, 0.1 M Na MES pH 6.5, 0.2 M Mg Acetate |
| **C2** | 30% MPD, 0.5 M Ammonium Sulfate, 0.1 M Na HEPES pH 7.5 |
| **C3** | 30% MPD, 0.1 M Na HEPES pH 7.5, 0.2 M Na Citrate |
| **C4** | 30% MPD, 5% PEG 4000, 0.1 M Na HEPES pH 7.5 |
| **C5** | 30% MPD, 10% PEG 4000, 0.1 M Imidazole HCl pH 8.0 |
| **C6** | 30% MPD, 20% Ethanol |
| **D1** | 35% MPD |
| **D2** | 35% MPD, 0.1 M Imidazole HCl pH 8.0 |
| **D3** | 40% MPD, 0.1 M TRIS HCl pH 8.5 |
| **D4** | 47% MPD, 0.1 M Na HEPES pH 7.5 |
| **D5** | 47% MPD, 2% *tert.*-Butanol |
| **D6** | 50% MPD |

JENA BIOSCIENCE
JenaBioscience GmbH
Loebstedter Str. 78
07749 Jena, Germany
Tel.: +49-3641-46 49 52
Fax: +49-3641-46 49 91
e-Mail: info@jenabioscience.com

# JBScreen 8
## (MPD/Alcohole based)

***Formulations***
**Cat.-No.: CS-108**

| | |
|---|---|
| **A1** | 50% MPD, 15% Ethanol, 0.01 M Na Acetate |
| **A2** | 50% MPD, 20% *iso*-Propanol, 0.05 M Na Acetate, 0.05 M Na Chloride |
| **A3** | 50% MPD, 0.1 M TRIS HCl pH 8.5, 0.2 M Ammonium Dihydrogen Phosphate |
| **A4** | 55% MPD |
| **A5** | 60% MPD, 0.1 M Na Acetate pH 4.6, 0.01 M Ca Chloride |
| **A6** | 60% MPD, 0.02 M Na Acetate |
| **B1** | 70% MPD, 0.1 M Na MES pH 6.5 |
| **B2** | 70% MPD, 0.1 M TRIS HCl pH 8.5 |
| **B3** | 20% Methanol, 0.1 M TRIS HCl pH 8.5, 0.01 M Ca Chloride |
| **B4** | 2% Ethanol, 0.1 M TRIS HCl pH 8.5 |
| **B5** | 5% Ethanol, 5% MPD, 0.1 M Na HEPES pH 7.5 |
| **B6** | 5% Ethanol, 5% MPD, 0.1 M TRIS HCl pH 8.5, 0.2 M Na Chloride |
| **C1** | 10% Ethanol, 0.1 M TRIS HCl pH 8.5 |
| **C2** | 12% Ethanol, 4% PEG 400, 0.1 M Na Acetate pH 4.6 |
| **C3** | 14% Ethanol, 5% Glycerol, 0.1 M TRIS HCl pH 8.5 |
| **C4** | 18% Ethanol, 0.1 M TRIS HCl pH 8.5 |
| **C5** | 20% Ethanol |
| **C6** | 20% Ethanol, 10% Glycerol |
| **D1** | 30% Ethanol, 10% PEG 6000, 0.1 M Na Acetate |
| **D2** | 45% Ethanol |
| **D3** | 50% Ethanol, 0.01 M Na Acetate |
| **D4** | 60% Ethanol, 1.5% PEG 6000, 0.05 M Na Acetate |
| **D5** | 60% Ethanol, 0.1 M Na Chloride |
| **D6** | 2% *iso*-Propanol, 0.1 M TRIS HCl pH 8.5, 0.01 M Mg Sulfate |

JENA BIOSCIENCE
JenaBioscience GmbH
Loebstedter Str. 78
07749 Jena, Germany
Tel.: +49-3641-46 49 52
Fax: +49-3641-46 49 91
e-Mail: info@jenabioscience.com

# JBScreen 9
## (Alcohole/Salt based)

## *Formulations*
**Cat.-No.: CS-109**

| | |
|---|---|
| **A1** | 5% *iso*-Propanol, 0.1 M Na HEPES pH 7.5 |
| **A2** | 10% *iso*-Propanol, 0.1 M Na Acetate pH 4.6, 0.2 M Ca Chloride |
| **A3** | 10% *iso*-Propanol, 0.1 M Na HEPES pH 7.5, 0.2 M Na Citrate |
| **A4** | 10% *iso*-Propanol, 0.1 M TRIS HCl pH 8.5, 0.01 M Mg Chloride |
| **A5** | 12% *iso*-Propanol, 0.1 M TRIS HCl pH 8.5, 0.05 Na Chloride |
| **A6** | 15% *iso*-Propanol, 0.1 M Na MES pH 6.5, 0.2 M Na Citrate |
| **B1** | 15% *iso*-Propanol, 0.1 M Na HEPES pH 7.5, 0.2 M Na Citrate |
| **B2** | 15% *iso*-Propanol, 0.1 M Na HEPES pH 7.5, 0.2 M Mg Chloride |
| **B3** | 15% *iso*-Propanol, 0.1 M TRIS HCl pH 8.5, 0.2 M Ammonium Acetate |
| **B4** | 20% *iso*-Propanol, 0.1 M Na Acetate pH 4.6, 0.2 M Ca Chloride |
| **B5** | 20% *iso*-Propanol, 0.1 M Na HEPES pH 7.5, 0.2 M Na Citrate |
| **B6** | 25% *iso*-Propanol, 0.1 M Na HEPES pH 7.5, 0.1 M Mg Chloride |
| **C1** | 30% *iso*-Propanol, 0.1 M Na MES pH 6.5, 0.2 M Na Citrate |
| **C2** | 30% *iso*-Propanol, 0.1 M Na HEPES pH 7.5, 0.2 M Mg Chloride |
| **C3** | 30% *iso*-Propanol, 0.1 M TRIS HCl pH 8.5, 0.2 M Ammonium Acetate |
| **C4** | 25% *tert.*-Butanol, 0.1 M TRIS HCl pH 8.5, 0.1 M Ca Chloride |
| **C5** | 35% *tert.*-Butanol, 0.1 M Na Citrate pH 5.6 |
| **C6** | 0.2 M Ammonium Dihydrogen Phosphate |
| **D1** | 0.2 M K,Na Tartrate |
| **D2** | 0.2 M Mg Acetate |
| **D3** | 0.4 M Ammonium Dihydrogen Phosphate |
| **D4** | 0.4 M K,Na Tartrate |
| **D5** | 0.4 M K,Na Tartrate, 0.1 M TRIS HCl pH 8.5 |
| **D6** | 0.5 M Ammonium Dihydrogen Phosphate, 0.2 M Na Citrate |

JENA BIOSCIENCE
JenaBioscience GmbH
Loebstedter Str. 78
07749 Jena, Germany
Tel.: +49-3641-46 49 52
Fax: +49-3641-46 49 91
e-Mail: info@jenabioscience.com

## *Crystal Screen* — Reagent Formulation

| Tube Number | Salt | Buffer † | Precipitant |
|---|---|---|---|
| 1. | 0.02 M Calcium Chloride dihydrate | 0.1 M Sodium Acetate trihydrate pH 4.6 | 30% v/v 2-Methyl-2,4-pentanediol |
| 2. | None | None | 0.4 M Potassium Sodium Tartrate tetrahydrate |
| 3. | None | None | 0.4 M mono-Ammonium dihydrogen Phosphate |
| 4. | None | 0.1 M Tris Hydrochloride pH 8.5 | 2.0 M Ammonium Sulfate |
| 5. | 0.2 M tri-Sodium Citrate dihydrate | 0.1 M HEPES - Na pH 7.5 | 30% v/v 2-Methyl-2,4-pentanediol |
| 6. | 0.2 M Magnesium Chloride hexahydrate | 0.1 M Tris Hydrochloride pH 8.5 | 30% w/v Polyethylene Glycol 4000 |
| 7. | None | 0.1 M Sodium Cacodylate pH 6.5 | 1.4 M Sodium Acetate trihydrate |
| 8. | 0.2 M tri-Sodium Citrate dihydrate | 0.1 M Sodium Cacodylate pH 6.5 | 30% v/v iso-Propanol |
| 9. | 0.2 M Ammonium Acetate | 0.1 M tri-Sodium Citrate dihydrate pH 5.6 | 30% w/v Polyethylene Glycol 4000 |
| 10. | 0.2 M Ammonium Acetate | 0.1 M Sodium Acetate trihydrate pH 4.6 | 30% w/v Polyethylene Glycol 4000 |
| 11. | None | 0.1 M tri-Sodium Citrate dihydrate pH 5.6 | 1.0 M mono-Ammonium dihydrogen Phosphate |
| 12. | 0.2 M Magnesium Chloride hexahydrate | 0.1 M HEPES - Na pH 7.5 | 30% v/v iso-Propanol |
| 13. | 0.2 M tri-Sodium Citrate dihydrate | 0.1 M Tris Hydrochloride pH 8.5 | 30% v/v Polyethylene Glycol 400 |
| 14. | 0.2 M Calcium Chloride dihydrate | 0.1 M HEPES - Na pH 7.5 | 28% v/v Polyethylene Glycol 400 |
| 15. | 0.2 M Ammonium Sulfate | 0.1 M Sodium Cacodylate pH 6.5 | 30% w/v Polyethylene Glycol 8000 |
| 16. | None | 0.1 M HEPES - Na pH 7.5 | 1.5 M Lithium Sulfate monohydrate |
| 17. | 0.2 M Lithium Sulfate monohydrate | 0.1 M Tris Hydrochloride pH 8.5 | 30% Polyethylene Glycol 4000 |
| 18. | 0.2 M Magnesium Acetate tetrahydrate | 0.1 M Sodium Cacodylate pH 6.5 | 20% Polyethylene Glycol 8000 |
| 19. | 0.2 M Ammonium Acetate | 0.1 M Tris Hydrochloride pH 8.5 | 30% v/v iso-Propanol |
| 20. | 0.2 M Ammonium Sulfate | 0.1 M Sodium Acetate trihydrate pH 4.6 | 25% w/v Polyethylene Glycol 4000 |
| 21. | 0.2 M Magnesium Acetate tetrahydrate | 0.1 M Sodium Cacodylate pH 6.5 | 30% v/v 2-Methyl-2,4-pentanediol |
| 22. | 0.2 M Sodium Acetate trihydrate | 0.1 M Tris Hydrochloride pH 8.5 | 30% w/v Polyethylene Glycol 4000 |
| 23. | 0.2 M Magnesium chloride hexahydrate | 0.1 M HEPES - Na pH 7.5 | 30% v/v Polyethylene Glycol 400 |
| 24. | 0.2 M Calcium Chloride dihydrate | 0.1 M Sodium Acetate trihydrate pH 4.6 | 20% v/v iso-Propanol |
| 25. | None | 0.1 M Imidazole pH 6.5 | 1.0 M Sodium Acetate trihydrate |
| 26. | 0.2 M Ammonium Acetate | 0.1 M tri-Sodium Citrate dihydrate pH 5.6 | 30 % v/v 2-Methyl-2,4-pentanediol |
| 27. | 0.2 M tri-Sodium Citrate dihydrate | 0.1 M HEPES - Na pH 7.5 | 20% v/v iso-Propanol |
| 28. | 0.2 M Sodium Acetate trihydrate | 0.1 M Sodium Cacodylate pH 6.5 | 30% w/v Polyethylene Glycol 8000 |
| 29. | None | 0.1 M HEPES - Na pH 7.5 | 0.8 M Potassium Sodium Tartrate tetrahydrate |
| 30. | 0.2 M Ammonium Sulfate | None | 30% w/v Polyethylene Glycol 8000 |
| 31. | 0.2 M Ammonium Sulfate | None | 30% w/v Polyethylene Glycol 4000 |
| 32. | None | None | 2.0 M Ammonium Sulfate |
| 33. | None | None | 4.0 M Sodium Formate |
| 34. | None | 0.1 M Sodium Acetate trihydrate pH 4.6 | 2.0 M Sodium Formate |
| 35. | None | 0.1 M HEPES - Na pH 7.5 | 0.8 M mono-Sodium dihydrogen phosphate<br>0.8 M mono-Potassium dihydrogen phosphate |
| 36. | None | 0.1 M Tris Hydrochloride pH 8.5 | 8% w/v Polyethylene Glycol 8000 |
| 37. | None | 0.1 M Sodium Acetate trihydrate pH 4.6 | 8% w/v Polyethylene Glycol 4000 |
| 38. | None | 0.1 M HEPES - Na pH 7.5 | 1.4 M tri-Sodium Citrate dihydrate |
| 39. | None | 0.1 M HEPES - Na pH 7.5 | 2% v/v Polyethylene Glycol 400, 2.0 M Ammonium Sulfate |
| 40. | None | 0.1 M tri-Sodium Citrate dihydrate pH 5.6 | 20% v/v iso-Propanol, 20% w/v Polyethylene Glycol 4000 |
| 41. | None | 0.1 M HEPES - Na pH 7.5 | 10% v/v iso-Propanol, 20% w/v Polyethylene Glycol 4000 |
| 42. | 0.05 M mono-Potassium dihydrogen Phosphate | None | 20% w/v Polyethylene Glycol 8000 |
| 43. | None | None | 30% w/v Polyethylene Glycol 1500 |
| 44. | None | None | 0.2 M Magnesium Formate |
| 45. | 0.2 M Zinc Acetate dihydrate | 0.1 M Sodium Cacodylate pH 6.5 | 18% w/v Polyethylene Glycol 8000 |
| 46. | 0.2 M Calcium Acetate hydrate | 0.1 M Sodium Cacodylate pH 6.5 | 18% w/v Polyethylene Glycol 8000 |
| 47. | None | 0.1 M Sodium Acetate trihydrate pH 4.6 | 2.0 M Ammonium Sulfate |
| 48. | None | 0.1 M Tris Hydrochloride pH 8.5 | 2.0 M mono-Ammonium dihydrogen Phosphate |
| 49. | 1.0 M Lithium Sulfate monohydrate | None | 2% w/v Polyethylene Glycol 8000 |
| 50. | 0.5 M Lithium Sulfate monohydrate | None | 15% w/v Polyethylene Glycol 8000 |

† *Buffer pH is that of a 1.0 M stock prior to dilution with other reagent components. pH with HCl or NaOH.*

*Crystal Screen contains fifty unique reagents. To determine the formulation of each reagent, simply read across the page.*

HAMPTON RESEARCH

*Solutions for Crystal Growth*

34 Journey
Aliso Viejo, CA 92656-3317 U.S.A.
Tel: (949) 425-1321 • Fax: (949) 425-1611
e-mail: tech@hrmail.com
Website: www.hamptonresearch.com

# APPENDIX THREE
# CLOSTRIDIAL NEUROTOXINS

**BOTULINUM AND TETANUS NEUROTOXINS: STRUCTURE, FUNCTION AND THERAPEUTIC UTILITY**

TURTON, K., CHADDOCK, J.A. & ACHARYA, K.R. (2002)

TRENDS BIOCHEM SCI, **27**, 552-558

# Botulinum and tetanus neurotoxins: structure, function and therapeutic utility

Kathryn Turton, John A. Chaddock and K. Ravi Acharya

**The toxic products of the anaerobic bacteria *Clostridium botulinum, Clostridium butyricum, Clostridium barati* and *Clostridium tetani* are the causative agents of botulism and tetanus. The ability of botulinum neurotoxins to disrupt neurotransmission, often for prolonged periods, has been exploited for use in several medical applications and the toxins, as licensed pharmaceutical products, now represent the therapeutics of choice for the treatment for several neuromuscular conditions. Research into the structures and activities of botulinum and tetanus toxins has revealed features of these proteins that might be useful in the design of improved vaccines, effective inhibitors and novel biopharmaceuticals. Here, we discuss the relationships between structure, mechanism of action and therapeutic use.**

DOI: 10.1016/S0968-0004(02)02177-1

Botulism results from intoxication by botulinum neurotoxin (BoNT) and is characterized by descending flaccid paralysis as a result of inhibition of acetylcholine release at the neuromuscular junction. There are seven botulinum neurotoxin serotypes (A–G) produced by bacteria of the genus *Clostridium*, and one of the factors in recovery from botulism depends on the BoNT serotype involved. Together with tetanus neurotoxin (TeNT) produced by *Clostridium tetani*, the BoNTs make up the clostridial neurotoxin (CNT) family. TeNT exhibits a high degree of sequence and structural homology to the BoNTs, in particular to BoNT/B, and is the causative agent of tetanus, which is characterized by spastic paralysis.

Although differing in clinical manifestation, the fundamental mode of action – inhibition of neurotransmission – is common to all CNTs. Inhibition of neurotransmitter release by the CNTs is caused by the specific cleavage of a group of proteins integral to the exocytotic process, the SNARE proteins (soluble NSF-attachment protein receptors). Cleavage of one or more of the SNARE proteins leads to a block in the release of vesicular contents to the extracellular environment. Therefore, in the case of BoNT action on the motorneuron, release of acetylcholine is prevented. Neither the duration of action, nor the electrophysiology of the effects of BoNT in the neuromuscular junction are equivalent across the different serotypes [1].

This article reviews the most recent interpretations of BoNT structure and mechanism of action with respect to the therapeutic use of the toxins, and draws comparisons with the data on TeNT.

**Kathryn Turton**
**K. Ravi Acharya***
Dept of Biology and Biochemistry, University of Bath, Claverton Down, Bath, UK BA2 7AY.
*e-mail: K.R.Acharya@bath.ac.uk

**John A. Chaddock**
Centre for Applied Microbiology and Research, Porton Down, Salisbury, UK SP4 0JG.

## CNT structure

To understand the components of BoNT responsible for its mechanism of action, it is necessary to introduce the basic CNT structure. CNTs are synthesized as single-chain polypeptides of ~150 kDa and are subsequently cleaved to form di-chain molecules (Fig. 1), in which the light (LC) and heavy chains (HC) are linked by a single disulfide bond. The 50-kDa LC acts as a zinc-dependent endopeptidase. The heavy chain contains two functional domains, each of ~50 kDa. The N-terminal half ($H_N$) is the translocation domain, known to form ion channels in lipid bilayers, and the C-terminal half ($H_C$) is the ganglioside-binding domain, which has a key role in binding to the target cell membrane and internalization of the toxin molecules into cholinergic neurons. The three functional domains are structurally distinct and arranged in a linear fashion, such that there is no contact between the LC and $H_C$ domains. Overall, BoNTs and TeNT share ~35% sequence identity [2]. The BoNT catalytic LC domains share up to 36% sequence identity [2], and the LC domains of BoNT/B and TeNT have over 50% identity [3].

Crystal structures of BoNT/A [4] (Fig. 2a), BoNT/B [5] and the $H_C$ domain of TeNT [6] (Fig. 2b) have been determined. The ganglioside-binding ($H_C$) domain of BoNT/A is made up of two sub-domains of roughly equal size. Both subdomains are composed of mainly β-sheet topology, joined by an α-helix (Fig. 2b). $H_C$ tilts away from the central translocation ($H_N$) domain such that its surface loops are accessible for binding. The $H_N$ domain has two unique structural features: a pair of long (105 Å) amphipathic helices, which resemble the coiled-coils of some viral proteins, and a long loop, known as the 'translocation belt', that wraps around the LC. The function of the $H_N$ domain is believed to be formation of pores and translocation of the catalytic domain into the neuron cytosol. However, the details of this are poorly understood and the pore-forming segment of $H_N$ remains to be identified. The catalytic domain (LC) contains the catalytic zinc atom and a conserved zinc-binding HExxH motif characteristic of zinc proteases. The catalytic site is located in a deep cleft on the protein surface, accessible by a channel. The translocation belt of the $H_N$ domain occludes the active site in BoNT/A, preventing access to the zinc atom, but is

  PII: S0968-0004(02)02177-1

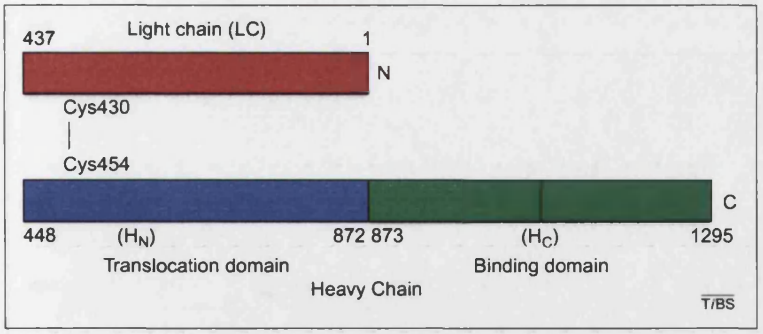



**Fig. 1.** The di-chain structure of a Clostridial neurotoxin – botulinum neurotoxin A (BoNT/A). Clostridial neurotoxins are ~150-kDa proteins, synthesized as single-chain polypeptides and post-translationally nicked to form di-chain molecules. They share the same domain architecture and overall structure. The light and heavy chains of BoNT/A are linked by a single disulfide bond, Cys430–Cys454. The light chain (LC), shown in red, functions as zinc-dependent endopeptidase and contains the catalytic zinc atom and HExxH motif associated with zinc-dependent proteases. The heavy chain comprises two functional domains of roughly equal size. The N-terminal section ($H_N$), shown in blue, is the translocation domain, which forms ion channels spanning endosomal membranes and is thought to be involved in translocation and activation of the LC. The C-terminal section ($H_C$), shown in green, is the binding domain.

less effective in BoNT/B partly because the belt is shorter in BoNT/B than BoNT/A.

Although the crystal structure of holo-TeNT has not yet been reported, similarities have been observed between the $H_C$ domains of BoNT serotypes A and B and that of TeNT (Fig. 2b). Like BoNT-$H_C$, the $H_C$ domain of TeNT is divided into two subdomains. It appears, from deletion mutagenesis studies [7] and structure-based sequence alignment [8], that the C-terminal subdomain is also involved in ganglioside binding. The jelly-roll topology of the N-terminal subdomain is similar to the structures of many lectins (carbohydrate-binding proteins) and might contain the ganglioside-binding motif. The crystal structures of TeNT in complex with a synthetic ganglioside analogue [9] and with carbohydrates [10] have provided insight into the ganglioside binding process, as discussed later.

## Mechanism of action

BoNT intoxication (summarized in Fig. 3) occurs through a multi-step process involving each of the toxin functional domains, and can be described as the outcome of three discrete stages.

### *Binding to target cell and internalization*

BoNTs bind to cholinergic nerve terminals by their $H_C$ domains and are subsequently internalized, possibly by receptor-mediated endocytosis. Binding is thought to require the presence of gangliosides (in particular GD1b, GT1b and GQ1b) [11]. However, as a result of ganglioside distribution studies and affinity binding studies, we know that gangliosides are not the sole molecule to facilitate BoNT binding. A 'double-receptor' model has been proposed [11] whereby BoNTs must also bind to a protein receptor before internalization can occur. The precise identity of the receptor for each member of the CNT family has remained elusive, although there is some evidence to suggest that synaptotagmin could be involved in the internalization of serotypes A, B and E [12–14]. Internalization of TeNT is known to be protease-sensitive and the requirement for lipid rafts and one or more GPI-anchored protein has now been confirmed [15] to be important in the TeNT-intoxication process. Once bound to the surface of the neuron, CNTs are internalized into acidic compartments through an endocytic process that is temperature and energy-dependent. At this stage, the toxins can no longer be neutralized by antisera.

### *Translocation*

In common with several other bacterial toxins, it is proposed that there is a pH-dependent structural rearrangement of the toxin inside an acidic compartment within the cell [16]. The identity of the intracellular compartment from which BoNTs translocate has not been categorically determined, although recycling synaptic vesicles and the endosomal compartment are obvious candidates. The reduction in pH is believed to trigger a structural change in BoNT, leading to greater hydrophobicity in the molecule and facilitating penetration of the lipid bilayer. It has been shown that BoNT serotypes A and E form ion channels in phospholipid bilayers [17] and PC12 membranes [18] under conditions similar to those believed to exist *in vivo*. Active LC is translocated into the cytosol, where it interacts with, and subsequently cleaves, SNAREs.

### *Inhibition of neurotransmitter release*

SNARE proteins are involved in the fusion of synaptic vesicles with the plasma membrane [19] and thus the action of BoNT-LC is to prevent exocytosis. At a more specific level, cleavage of SNARE proteins by BoNT inhibits the release of acetylcholine (ACh) at the neuromuscular junction, leading to inhibition of neurotransmission. It has been shown that cleavage of individual SNARE proteins does not prevent SNARE complex formation, but results in a non-functional complex where the coupling between $Ca^{2+}$ influx and fusion is disrupted (see [20] for discussion). The role of $Ca^{2+}$ is fundamental to the process of CNT-dependent inhibition of neurotransmitter release as increasing the $Ca^{2+}$ concentration in the synaptic terminal partially reverses the effect of BoNT/A [21].

TeNT binds to gangliosides on the motorneuron surface in a similar manner to the BoNTs but, once internalized, TeNT undergoes retrograde axonal transport to the central nervous system, transcytosing across the synapse to internalize into the interneurons that regulate motorneuron activity. The spastic paralysis that results from intoxication with TeNT arises from the loss of spinal inhibitory control of motorneuron activity [22,23].

## Structural basis of BoNT activity

### *Binding and internalization*

Data obtained for the structure of the $H_C$ domains of BoNT/A [4], BoNT/B [5] and TeNT [6] reveals that the

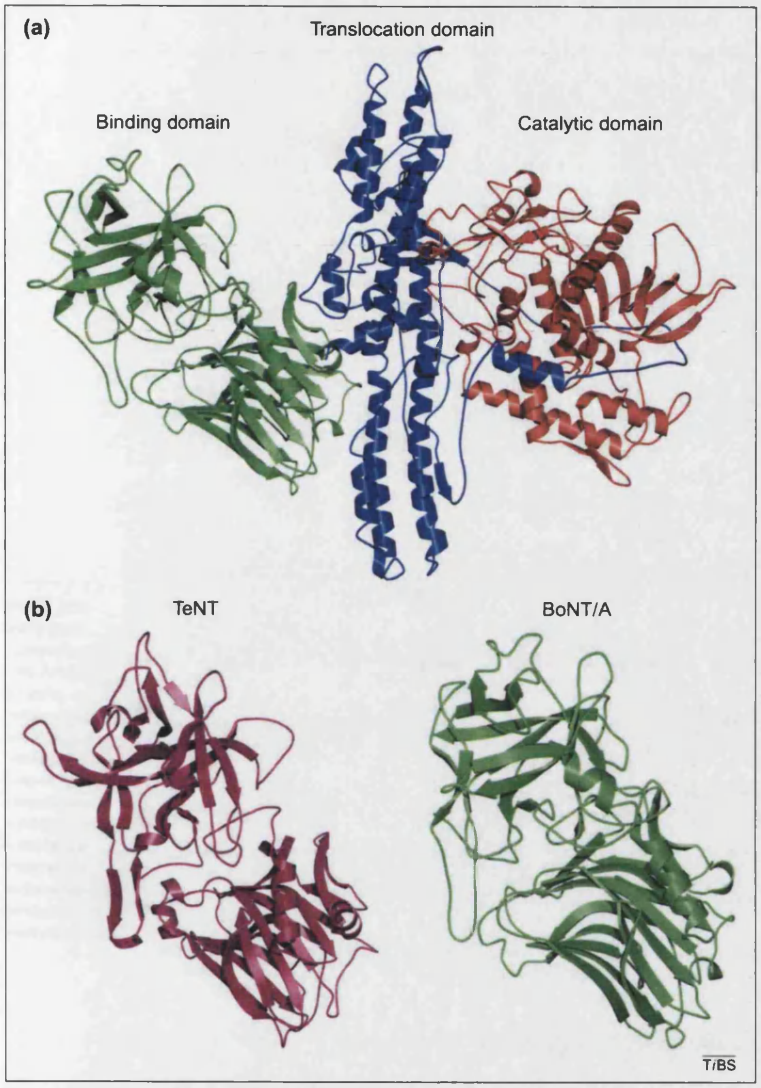



**Fig. 2.** (a) Crystal structure of botulinum neurotoxin A (BoNT/A). The three functional domains are structurally distinct and oriented such that there are no contacts between the catalytic and binding domains. The binding domain is made up of two subdomains of roughly equal size: the N-terminal subdomain comprises a jelly roll motif while the C-terminal subdomain (which is believed to contain the ganglioside-binding motif) adopts a β-trefoil fold. The central translocation domain has two unique structural features – a pair of long helices that are reminiscent of a viral coiled-coil motif and a long loop (translocation belt) that wraps around the catalytic domain and occludes the active site. The catalytic domain houses the catalytic zinc atom and the conserved HExxH motif characteristic of zinc-dependent endopeptidases. (b) Botulinum neurotoxin A and tetanus neurotoxin (TeNT) binding domains. Structural comparisons drawn between the binding domains ($H_C$) of BoNT/A and TeNT reveal their overall architecture is similar. In both toxins the binding domain comprises two subdomains, the N-terminal of which has a jelly-roll, lectin-like fold and the C-terminal of which adopts a β-trefoil fold. The subdomains are linked by a single α-helix. The C-terminal subdomain contains the ganglioside-binding site and the postulated five-residue ganglioside-binding motif.

binding domain is composed of two sub-domains. The N-terminal half adopts a jelly-roll, lectin-like fold and the C-terminal half has a β-trefoil fold. The binding domain binds to gangliosides on the surface of motorneurons. As described previously, BoNTs bind to gangliosides of the GT1b and GD1b classes with low affinity [24]. Gangliosides consist of sialic acid-containing oligosaccharide linked to ceramide.

Sequence-to-structure alignment of the CNTs reveals a set of five conserved residues (in BoNT/A, these are Glu1202, Ser1263, Trp1265, Tyr1266 and Gly1278) within the C-terminal subdomain of $H_C$ that are postulated to form a ganglioside-binding motif [8]. Neutralizing antibody data have also implicated the stretch of amino acids 1266–1272 in BoNT/A in ganglioside binding [10]. This motif lies close to a shallow cleft on the surface of BoNT/A and could accommodate the proposed protein co-receptor.

Since initial structural analysis of TeNT-$H_C$, binding studies and cocrystallisation experiments have been performed on this fragment [9,10,25] that have revealed clues to the mechanism of toxin internalization into cholinergic neurons. The five residues of the proposed CNT ganglioside binding motif (Trp1289, His1271, His1293, Tyr1290 and Phe1218) have been implicated in ganglioside-binding in TeNT [9] and Tyr1290 has been shown, by site-directed mutagenesis, to be a key residue in ganglioside binding [7]. In addition, Arg1226, Asn1216, Asp1214, Asp1147 and Tyr1229 have been shown to line pockets on the surface of TeNT where gangliosides bind. Binding studies with TeNT have identified a carbohydrate moiety essential for TeNT (or TeNT-$H_C$) binding, and crystallographic studies have revealed two carbohydrate-binding areas on the surface on the C-terminal domain of TeNT-$H_C$ [10]. These two binding sites can interact with two separate ganglioside molecules, hence crosslinking could occur on the motorneuron surface. The crystal structure of TeNT-$H_C$ bound to lactose reveals that the lactose binding site appears to be composed of residues of the proposed CNT ganglioside-binding motif and that the lactose residue is bound in an orientation consistent with that for ganglioside-binding [10]. Superposition of the $H_C$ C-terminal sub-domains of BoNT/A, BoNT/B and TeNT reveals that they have similar binding clefts to accommodate sialic acid [5]. This has led to the proposition of a common model of interaction of GT1b gangliosides with these toxins.

CNTs are structurally homologous with respect to their $H_C$ domains but sequence identity is low, which could be important in determining the specificity of each toxin for its protein co-receptor. Structural features of BoNT/B in complex with sialyllactose [5] are similar to that of apo-BoNT/B, suggesting a 'lock-and-key' mechanism of binding between neurotoxin and ganglioside. This observation supports the idea that ganglioside-binding brings BoNTs into close proximity with a protein receptor, facilitating toxin recognition and internalization.

Once the CNT has bound to its receptor(s) on the motorneuron, it is internalized inside a vesicle, which the LC must ultimately leave to exert effects on the nervous system. In the case of BoNTs, exit from the vesicular compartment or endosome is the only means of accessing cytosolic protein targets (the SNARE proteins). By contrast, TeNT must be

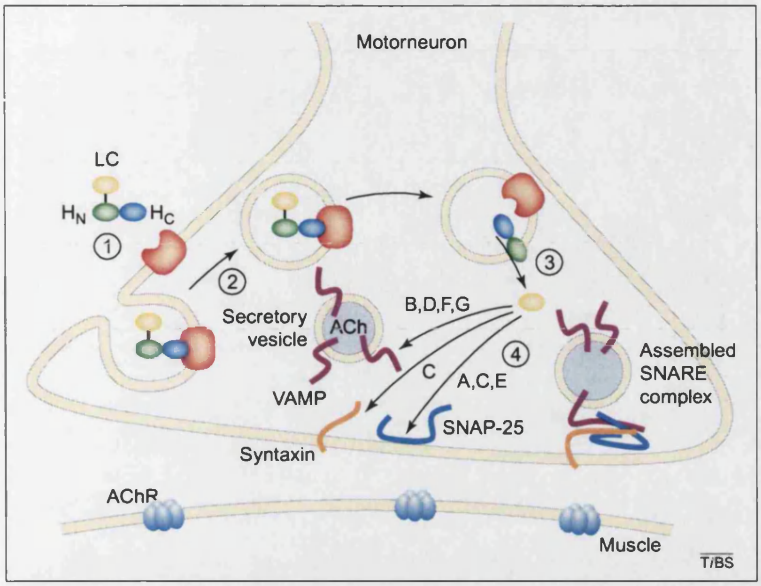



**Fig. 3.** Mechanism of action of botulinum neurotoxin (BoNT). Each of the structural elements of botulinum neurotoxin, the light chain (LC) endopeptidase (yellow), the $H_N$ translocation domain (green) and the $H_C$ cell-binding domain (blue) have a role in the mechanism of action of neurotoxin. The first stage in the intoxication process (interaction with gangliosides and an, as yet unidentified, protein receptor) (1) is followed by stage two, internalization of the toxin-receptor complex into an intracellular vesicle (2). The third stage (translocation) is characterized by release of the light chain endopeptidase from an acidic intracellular compartment into the cytosol (3). Once liberated from the vesicle, the light chain performs the final stage of intoxication; highly specific proteolytic cleavage of one of the proteins of the SNARE complex (4). BoNTs B, D, F and G cleave proteins of the VAMP family (purple) and BoNTs A, C and E cleave SNAP-25 (light blue). BoNT/C also has the capacity to cleave syntaxin (orange). Cleaved SNARE proteins are competent for facilitating docking of the secretory vesicle with the synaptic membrane, but fusion of the vesicle is compromised. Thus neurotransmitter release is inhibited. Abbreviations: $H_N$, heavy chain N-terminal subdomain; $H_C$, heavy chain C-terminal subdomain; SNARE, soluble NSF-attachment protein receptors; VAMP, vesicle-associated membrane protein; SNAP-25, synaptosomal-associated protein of 25 kDa; ACh, acetylcholine; AChR, acetylcholine receptor.

retrogradely transported to the motorneuron cell body, from which it can access the inhibitory neurons of the spinal cord.

*Pore formation*

A candidate transmembrane region identified in BoNT/A increases the permeability of lipid bilayers [4]. Analysis of the crystal structure reveals that this region does not form a helix at neutral pH, although it may undergo structural rearrangement at low (i.e. endosomal) pH. The BoNT/A residues most likely to titrate at low pH, His551 and His560, are located in a loop connecting the translocation belt to the main body of the $H_N$ domain, which might rearrange exposing hydrophobic residues. There are several clusters of negative charge on the surface of the $H_N$ domain that could be involved in belt movement and which could contribute to the structural changes that lead to pore formation [2].

BoNT/A and BoNT/B $H_N$ domains undergo conformational changes at low pH values, culminating in the formation of ion channels spanning artificial membranes, and the activity of the channels formed is maximal at endosomal pH [17]. Channels formed by BoNT/C in lipid membranes are similar to those formed by diphtheria toxin and also only form at low pH [26]. BoNT channels are specific for cations (although not as highly specific as voltage-gated ion channels) but can be blocked *in vitro* by chloroquine.

However, it is not yet clear whether the channels formed by BoNTs are also the route by which the toxin LC enters the cytosol, because of the small pore size of the ion channels studied [18]. Data indicating that there is a dramatic change in LC structure at low pH [27] suggest that these conformational changes may lead to creation of an alternative LC structure that is more suited to translocation. Although this has not been determined *in vivo*, the structural changes induced in BoNT/A-LC by low pH are completely reversible [27].

Unfortunately, pore formation is currently the least well understood step in the mechanism of BoNT action and further investigation is required to uncover the extent and implication of the structural changes that occur at low pH.

*LC release and neurotoxin activation*

For BoNT to exert its effect on neurotransmission requires the activation of the inactive single-chain 150-kDa toxin by proteolytic nicking within a surface-exposed loop at the LC-$H_N$ junction. Several bacterial and tissue proteases are able to perform this cleavage reaction to generate the active di-chain neurotoxin. Following nicking, the LC and HC remain associated through non-covalent interactions and a single disulfide bond. Reduction of this interchain disulfide bond (Cys430–Cys454 in BoNT/A) is the second stage of LC activation, and is necessary to render the LC free to enter the motorneuron cytosol and interact with substrate. As mentioned previously, there is a dramatic change in LC structure at low pH [27]. Such structural changes could contribute to the rearrangement of the translocation belt (if indeed the $H_N$ domain is still in close proximity to the LC at the point of translocation) and lead to final activation of the LC.

Interestingly, both the HC and LC of TeNT and BoNT serotypes A, B and E are phosphorylated *in vitro* by Src, leading to a marked increase in catalytic activity [28] and increased structural stability [29]. It has also been demonstrated that the more phosphorylation sites that are phosphorylated, the greater the increase in SNARE cleavage. These data indicate that neuronal signaling pathways could be intimately connected with BoNT-LC activity. It is, therefore, possible that completion of LC activation ultimately occurs in the cytosol following translocation from the low-pH compartment.

*SNARE-protein cleavage and the catalytic mechanism*

BoNTs disrupt neurotransmission by cleaving SNARE proteins in a highly specific manner, no two serotypes cleave the same SNARE protein at the same peptide bond. In general terms, BoNTs A and E

act on SNAP-25 (synaptosomal-associated protein of 25 kDa); BoNTs B, D, F and G cleave synaptobrevin (also known as vesicle-associated membrane protein, or VAMP); and BoNT $C_1$ can act on both SNAP-25 and syntaxin. However, it has recently become clear that some SNARE proteins are insensitive to cleavage by specific BoNTs and that the above general statement can be misleading (reviewed in Ref. [20]). For example, the presence of TeNT-insensitive VAMP7 (TI-VAMP) and BoNT B-, D-, F- and G-insensitive VAMP8 (endobrevin) [30] and SNAP-23 in a range of cells and their proposed roles in endocytosis and exocytosis does not support the simple interpretation of the relationship between SNAREs and CNTs. In several cases, for example human SNAP-23, resistance to cleavage by BoNT is determined by a single amino acid substitution [31].

The specificity of BoNTs for their SNARE-protein substrate has been suggested to be because of their ability to recognize and interact with a nine-residue motif (the SNARE secondary recognition motif; SSR motif) within the SNARE protein. Indeed, there is evidence [31–34] for a requirement for all the BoNT LC domains to interact with one or more SSR motifs distant from the site of endopeptidase activity. The SSR motif occurs twice in synaptobrevin, twice in syntaxin and at four positions in SNAP-25, and these motifs are predicted to form amphipathic α-helices. The SSR motif is conserved throughout several species, which suggests that it is probably an essential part of SNARE function. Peptides corresponding to the nine-residue SSR motif have been shown to inhibit CNT action [32], indicating that these motifs are involved in substrate recognition by CNTs despite being distinct from the cleavage sites of the different toxins. The crystal structure of BoNT/B in complex with synaptobrevin 2 [35] also reveals the involvement of surface loops on the LC domain in binding the SSR motif at a site distal to the catalytic site.

The BoNT active site is buried (~20 Å deep) in the LC, has negative surface charge and is accessible by a channel. The CNT active site contains a single zinc atom which has historically been proposed to have a functional, but not structural role, as the zinc can be removed by the use of metal chelators but reacquired by addition of free zinc without significant effects on the secondary structure [36]. However, recent studies investigating the potential structural role of the zinc atom in BoNT activity have been inconclusive [37–39]. TeNT would appear to be less restricted as the active site zinc atom can be exchanged with other divalent transition metals to form active toxin [40].

Residues His 222, Glu223 and His226 of BoNT/A form part of the conserved HExxH motif characteristic of zinc proteases and contribute directly to the reaction mechanism through co-ordination of the active site zinc (His 222 and His226) and a water molecule (Glu223). Glu261 is the third ligand to co-ordinate the zinc [4,41]. The catalytic mechanism of BoNT/A is proposed to be similar to that of thermolysin [42] based on structural and sequence similarities between their active sites. Mutation of Glu223 to Asp leads to almost total loss of activity [43] and removal of His226 abolishes the catalytic activity of BoNT/A [44]. Analysis of the structure of BoNT/B has confirmed that the key residues involved in the mechanism of action are identical to those observed in BoNT/A. The crystal structure of BoNT/B bound to synaptobrevin 2 reveals that residues Arg369 and Tyr372 are located in the proximity of the peptide bond cleaved by the toxin [35]. These residues are strictly conserved throughout all CNTs, and the equivalent residues in TeNT (Glu271 and Tyr375) have been shown, by site-directed mutagenesis, to be involved in protease activity [45]. Mutation of the corresponding residues in BoNT/A (Arg362 and Tyr365 to alanine and phenylalanine, respectively) dramatically reduces $k_{cat}$, but does not affect zinc binding or substrate binding [46], and it has been proposed that these residues are involved in transition state stabilization.

At the presynaptic nerve terminal, the core of the SNARE complex is stabilized by a bundle of four tightly packed helices [47], one from SNAP-25, one from syntaxin and two from synaptobrevin. The hydrophobic residues within these helices are conserved in the SNARE proteins and are known as the SNARE motif (this is distinct from the SSR motif mentioned previously). The SNARE complex is extremely stable and must be actively disassembled following neurotransmitter release. The structure of BoNT/B in complex with synaptobrevin 2 [35] revealed that the SNARE motif fits into the active site in a random coil conformation. In the SNARE complex, this motif is helical and cannot enter the active site. This explains why BoNTs can only act on free SNAREs and suggests a role for BoNTs in mopping up free SNAREs in solution (see Ref. [48] for a discussion on the action of clostridial neurotoxins on SNARE assembly).

### *Duration of activity*

The extended duration of BoNT effects at the neuromuscular junction is one of the key features that has driven the widespread use of BoNT as a therapeutic agent. However, there are significant differences in the *in vivo* duration of effect between the different BoNTs. This has been most fully studied through a comparison of the effects of BoNT/A and BoNT/E. BoNT/A induces long-term inhibition (months) whereas recovery from BoNT/E effects is rapid (weeks) [1]. The debate as to the reason for the disparity between BoNTs has centered around two hypotheses: first, that the BoNT/A-truncated SNAP-25 product persists in the SNARE complex and interferes with neurotransmitter release; second, there is a difference in the life-time of the LC in the cytosol, with BoNT/A persisting for longer than BoNT/E. These two hypotheses are not mutually exclusive, and recent findings support the view that

both persistence of the LC and effects of the truncated product have a role [49]. However, it is important to note that the biochemistry of BoNT action is not the only factor that determines longevity of effect. There is a significant contribution from the biology of the system in which the rate, frequency and success of nerve sprouting have a role to play in determining the rate of recovery (reviewed in Ref. [21]).

### Medical applications of BoNTs

Despite being extremely poisonous to humans, BoNTs are highly effective therapeutic agents and accomplished research tools. It was observed at the beginning of the 20th century that, by injection of botulinum toxin, localized paralysis of individual muscle groups could be achieved without causing botulism in the recipient (reviewed in Ref. [50]). Preparations of BoNT/A have been approved by the US Food and Drug Administration as a biopharmaceutical for use in treating strabismus, blepharospasm and hemifacial spasm. However, the use of BoNT has now extended to cover a wide variety of disorders [51], including those that do not have a neuromuscular basis. For example, the effects of BoNT on the autonomic nervous system have resulted in its use as an effective treatment for axillary hyperhidrosis [52] (sweating), and some success has been reported for the treatment of myofascial pain [53] and tension and migraine headaches [54].

However, as with any therapeutic agent, there are some side effects associated with the use of BoNT/A [50,51] such as dysphagia (following injections into the neck muscles), although weakening of surrounding muscles is rare. Inherent in the use of a protein-based therapeutic is the potential for antibody formation leading to a decrease in effectiveness of the treatment. Such secondary non-responders are seen in a relatively low percentage of patients, most commonly with patients requiring large doses of BoNT, often on repeated occasions. Patients immune to BoNT/A might benefit from injection with serotypes B and F, although this has yet to be proven and, as discussed previously, the duration of effect of each serotype differs. BoNT/C, shown to induce paralysis as long-lasting as that caused by BoNT/A [55], might be developed into a valid alternative therapy. Since the majority of the immune response is generated toward the $H_C$ fragment, future protein engineering of hybrid toxins could provide one route to prolong the therapeutic efficacy of BoNT treatment.

### Future directions

Although some progress has been made in recent years, identification and characterization of the protein receptors for the BoNTs and determination of the mechanism of specificity of CNT binding domains for their receptors is an outstanding problem. Furthermore, understanding the mechanism of LC translocation and activation within the motorneuron, including the effects of pH on the tertiary structures of BoNTs, will be crucial for rational design of engineered BoNT therapeutics. Further structural studies on the endopeptidase domains of BoNTs and TeNT, including the structural basis behind BoNT substrate specificity, might lead to the development of serotype-specific inhibitors. It is clear that the headway made in the structural knowledge of BoNT and TeNT has already paid dividends in the understanding of these complex proteins, and it is essential that future assessment of the mechanism of BoNT action take into account functionally relevant domains of the toxins, in functionally relevant physiological conditions.

Use of fragments of BoNT for the therapeutics of the future is also exciting. For example, harnessing the properties of the BoNT LC endopeptidase fragments for the creation of a range of 'designer' therapeutics is a real possibility following the successful retargeting of the LC/A domain to cells of neuronal [56] and non-neuronal [57] origin. Additionally, the ability of BoNTs to transport large polypeptides across the membranes could be harnessed for the delivery of biopharmaceuticals to cytosolic targets [58]. TeNT conjugates have been used to transport DNA [59] and enzymes [60] into neurons, and derivatives of BoNT/A and BoNT/B can target compounds specifically to human neuroblastoma cells [61]. The bacterium *C. botulinum* itself might also be useful in treating cancer – apathogenic *Clostridia* specifically colonize hypoxic areas of tumors [62] and might be useful for transporting anti-cancer drugs specifically into tumor cells [63].

It can be seen that the botulinum neurotoxins are of great interest to the medical and scientific communities. Despite causing disease, they have become valuable research tools and have wide-ranging applications as pharmaceuticals. As the structure and mechanism of action of the toxins are further dissected, the development of vaccines, serotype-specific inhibitors and novel therapeutics will undoubtedly follow.

**Acknowledgements**
We apologize for the omission of many relevant citations imposed by space limitations. We wish to acknowledge the helpful comments of Bernard Poulain and Neil Isaacs during preparation of this manuscript. Our work on BoNTs is supported by a joint post-graduate studentship to K.T. between the University of Bath and CAMR, UK. K.R.A. wishes to acknowledge the Royal Society and the Leverhulme Trust, UK for a Senior Research Fellowship.

**References**

1 Eleopra, R. *et al.* (1998) Different time courses of recovery after poisoning with botulinum neurotoxin serotypes A and E in humans. *Neurosci. Lett.* 256, 135–138
2 Lacy, D.B. and Stevens, R.C. (1999) Sequence homology and structural analysis of the clostridial neurotoxins. *J. Mol. Biol.* 291, 1091–1104
3 Kurazono, H. *et al.* (1992) Minimal essential domains specifying toxicity of the light chains of tetanus toxin and botulinum neurotoxin type A. *J. Biol. Chem.* 267, 14721–14729
4 Lacy, D.B. *et al.* (1998) Crystal structure of botulinum neurotoxin type A and implications for toxicity. *Nat. Struct. Biol.* 5, 898–902
5 Swaminathan, S. and Eswaramoorthy, S. (2000) Structural analysis of the catalytic and binding sites of Clostridium botulinum neurotoxin B. *Nat. Struct. Biol.* 7, 693–699
6 Umland, T.C. *et al.* (1997) Structure of the receptor binding fragment HC of tetanus neurotoxin. *Nat. Struct. Biol.* 4, 788–792
7 Sutton, J.M. *et al.* (2001) Tyrosine-1290 of tetanus neurotoxin plays a key role in its binding to gangliosides and functional binding to neurones. *FEBS Lett.* 493, 45–49
8 Ginalski, K. *et al.* (2000) Structure-based sequence alignment for the beta-trefoil subdomain of the clostridial neurotoxin family

provides residue level information about the putative ganglioside binding site. *FEBS Lett.* 482, 119–124
9 Fotinou, C. *et al.* (2001) The crystal structure of tetanus toxin Hc fragment complexed with a synthetic GT1b analogue suggests cross-linking between ganglioside receptors and the toxin. *J. Biol. Chem.* 276, 32274–32281
10 Emsley, P. *et al.* (2000) The structures of the HC fragment of tetanus toxin with carbohydrate subunit complexes provide insight into ganglioside binding. *J. Biol. Chem.* 275, 8889–8894
11 Montecucco, C. (1986) How do tetanus and botulinum toxins bind to neuronal membranes? *Trends Biochem. Sci.* 11, 314–317
12 Li, L. and Singh, B.R. (1998) Isolation of synaptotagmin as a receptor for types A and E botulinum neurotoxin and analysis of their comparative binding using a new microtiter plate assay. *J. Nat. Toxins* 7, 215–226
13 Nishiki, T. *et al.* (1996) The high-affinity binding of Clostridium botulinum type B neurotoxin to synaptotagmin II associated with gangliosides GT1b/GD1a. *FEBS Lett.* 378, 253–257
14 Nishiki, T. *et al.* (1994) Identification of protein receptor for Clostridium botulinum type B neurotoxin in rat brain synaptosomes. *J. Biol. Chem.* 269, 10498–10503
15 Munro, P. *et al.* (2001) High sensitivity of mouse neuronal cells to tetanus toxin requires a GPI-anchored protein. *Biochem. Biophys. Res. Commun.* 289, 623–629
16 Montecucco, C. *et al.* (1994) Bacterial protein toxins penetrate cells via a four-step mechanism. *FEBS Lett.* 346, 92–98
17 Blaustein, R.O. *et al.* (1987) The N-terminal half of the heavy chain of botulinum type A neurotoxin forms channels in planar phospholipid bilayers. *FEBS Lett.* 226, 115–120
18 Sheridan, R.E. (1998) Gating and permeability of ion channels produced by botulinum toxin types A and E in PC12 cell membranes. *Toxicon* 36, 703–717
19 Bajjalieh, S.M. (1999) Synaptic vesicle docking and fusion. *Curr. Opin. Neurobiol.* 9, 321–328
20 Humeau, Y. *et al.* (2000) How botulinum and tetanus neurotoxins block neurotransmitter release. *Biochimie* 82, 427–446
21 Meunier, F.A. *et al.* (2002) Botulinum neurotoxins: from paralysis to recovery of functional neuromuscular transmission. *J. Physiol. Paris* 96, 105–113
22 Curtis, D.R. and De Groat, W.C. (1968) Tetanus toxin and spinal inhibition. *Brain Res.* 10, 208–212
23 Pellizzari, R. *et al.* (1999) Tetanus and botulinum neurotoxins: mechanism of action and therapeutic uses. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 354, 259–268
24 Halpern, J.L. and Neale, E.A. (1995) Neurospecific binding, internalization, and retrograde axonal transport. *Curr. Top. Microbiol. Immunol.* 195, 221–241
25 Herreros, J. *et al.* (2000) C-terminal half of tetanus toxin fragment C is sufficient for neuronal binding and interaction with a putative protein receptor. *Biochem. J.* 347, 199–204
26 Donovan, J.J. and Middlebrook, J.L. (1986) Ion-conducting channels produced by botulinum toxin in planar lipid membranes. *Biochemistry* 25, 2872–2876
27 Li, L. and Singh, B.R. (2000) Spectroscopic analysis of pH-induced changes in the molecular features of type A botulinum neurotoxin light chain. *Biochemistry* 39, 6466–6474
28 Ferrer-Montiel, A.V. *et al.* (1996) Tyrosine phosphorylation modulates the activity of clostridial neurotoxins. *J. Biol. Chem.* 271, 18322–18325
29 Encinar, J.A. *et al.* (1998) Structural stabilization of botulinum neurotoxins by tyrosine phosphorylation. *FEBS Lett.* 429, 78–82
30 Martinez-Arca, S. *et al.* (2000) Clostridial neurotoxin-insensitive vesicular SNAREs in exocytosis and endocytosis. *Biol. Cell* 92, 449–453
31 Vaidyanathan, V.V. *et al.* (1999) Proteolysis of SNAP-25 isoforms by botulinum neurotoxin types A, C, and E: domains and amino acid residues controlling the formation of enzyme-substrate complexes and cleavage. *J. Neurochem.* 72, 327–337
32 Pellizzari, R. *et al.* (1996) Structural determinants of the specificity for synaptic vesicle-associated membrane protein/synaptobrevin of tetanus and botulinum type B and G neurotoxins. *J. Biol. Chem.* 271, 20353–20358
33 Washbourne, P. *et al.* (1997) Botulinum neurotoxin types A and E require the SNARE motif in SNAP-25 for proteolysis. *FEBS Lett.* 418, 1–5
34 Rossetto, O. *et al.* (1994) SNARE motif and neurotoxin. *Nature* 372, 415–416
35 Hanson, M.A. and Stevens, R.C. (2000) Cocrystal structure of synaptobrevin-II bound to botulinum neurotoxin type B at 2.0 Å resolution. *Nat. Struct. Biol.* 7, 687–692
36 De Filippis, V. *et al.* (1995) Structural studies on the zinc-endopeptidase light chain of tetanus neurotoxin. *Eur. J. Biochem.* 229, 61–69
37 Fu, F.N. *et al.* (1998) Role of zinc in the structure and toxic activity of botulinum neurotoxin. *Biochemistry* 37, 5267–5278
38 Li, L. and Singh, B.R. (2000) Role of zinc binding in type A botulinum neurotoxin light chain's toxic structure. *Biochemistry* 39, 10581–10586
39 Simpson, L.L. *et al.* (2001) The role of zinc binding in the biological activity of botulinum toxin. *J. Biol. Chem.* 276, 27034–27041
40 Tonello, F. *et al.* (1997) Metal substitution of tetanus neurotoxin. *Biochem. J.* 322, 507–510
41 Rigoni, M. *et al.* (2001) Site-directed mutagenesis identifies active-site residues of the light chain of botulinum neurotoxin type A. *Biochem. Biophys. Res. Commun.* 288, 1231–1237
42 Matthews, B.W. (1988) Structural basis of the action of thermolysin and related zinc peptidases. *Acc. Chem. Res.* 21, 333–340
43 Li, L. *et al.* (2000) Probing the mechanistic role of glutamate residue in the zinc-binding motif of type A botulinum neurotoxin light chain. *Biochemistry* 39, 2399–2405
44 Zhou, L. *et al.* (1995) Expression and purification of the light chain of botulinum neurotoxin A: a single mutation abolishes its cleavage of SNAP-25 and neurotoxicity after reconstitution with the heavy chain. *Biochemistry* 34, 15175–15181
45 Rossetto, O. *et al.* (2001) Active-site mutagenesis of tetanus neurotoxin implicates TYR-375 and GLU-271 in metalloproteolytic activity. *Toxicon* 39, 1151–1159
46 Binz, T. *et al.* (2002) Arg362 and Tyr365 of the botulinum neurotoxin type A light chain are involved in transition state stabilization. *Biochemistry* 41, 1717–1723
47 Ossig, R. *et al.* (2000) Exocytosis requires asymmetry in the central layer of the SNARE complex. *EMBO J.* 19, 6000–6010
48 Hayashi, T. *et al.* (1994) Synaptic vesicle membrane fusion complex: action of clostridial neurotoxins on assembly. *EMBO J.* 13, 5051–5061
49 Keller, J.E. and Neale, E.A. (2001) The role of the synaptic protein SNAP-25 in the potency of botulinum neurotoxin type A. *J. Biol. Chem.* 276, 13476–13482
50 Johnson, E.A. (1999) Clostridial toxins as therapeutic agents: benefits of nature's most toxic proteins. *Annu. Rev. Microbiol.* 53, 551–575
51 Munchau, A. and Bhatia, K.P. (2000) Uses of botulinum toxin injection in medicine today. *BMJ* 320, 161–165
52 Heckmann, M. *et al.* (2001) Botulinum toxin A for axillary hyperhidrosis (excessive sweating). *N. Engl. J. Med.* 344, 488–493
53 Porta, M. (2000) A comparative trial of botulinum toxin type A and methylprednisolone for the treatment of myofascial pain syndrome and pain from chronic muscle spasm. *Pain* 85, 101–105
54 Silberstein, S. *et al.* (2000) Botulinum toxin type A as a migraine preventive treatment. *Headache* 40, 445–450
55 Eleopra, R. *et al.* (1997) Botulinum neurotoxin serotype C: a novel effective botulinum toxin therapy in human. *Neurosci. Lett.* 224, 91–94
56 Chaddock, J.A. *et al.* (2000) A conjugate composed of nerve growth factor coupled to a non-toxic derivative of Clostridium botulinum neurotoxin type A can inhibit neurotransmitter release *in vitro*. *Growth Factors* 18, 147–155
57 Chaddock, J.A. *et al.* (2000) Inhibition of vesicular secretion in both neuronal and nonneuronal cells by a retargeted endopeptidase derivative of Clostridium botulinum neurotoxin type A. *Infect. Immun.* 68, 2587–2593
58 Simpson, L.L. (2000) Identification of the characteristics that underlie botulinum toxin potency: Implications for designing novel drugs. *Biochimie* 82, 943–953
59 Knight, A. *et al.* (1999) Non-viral neuronal gene delivery mediated by the HC fragment of tetanus toxin. *Eur. J. Biochem.* 259, 762–769
60 Figueiredo, D.M. *et al.* (1997) Delivery of recombinant tetanus-superoxide dismutase proteins to central nervous system neurons by retrograde axonal transport. *Exp. Neurol.* 145, 546–554
61 Zdanovskaia, M.V. *et al.* (2000) Recombinant derivatives of clostridial neurotoxins as delivery vehicles for proteins and small organic molecules. *J. Protein Chem.* 19, 699–707
62 Lambin, P. *et al.* (1998) Colonisation of Clostridium in the body is restricted to hypoxic and necrotic areas of tumours. *Anaerobe* 4, 183–188
63 Minton, N.P. *et al.* (1995) Chemotherapeutic tumour targeting using clostridial spores. *FEMS Microbiol. Rev.* 17, 357–364

# APPENDIX FOUR
# FRAGMENTS OF BOTULINUM NEUROTOXIN A

**SEQUENCE ALIGNMENT OF CLOSTRIDIAL NEUROTOXINS**

**EXPRESSION AND PURIFICATION OF $LH_N/A$**

**CRYSTALLIZATION CONDITIONS REPORTED FOR CNTs**

**CRYSTALLIZATION CONDITIONS REPORTED FOR $Zn^{2+}$ PROTEASES**

**ANALYSIS OF CD SPECTRA**

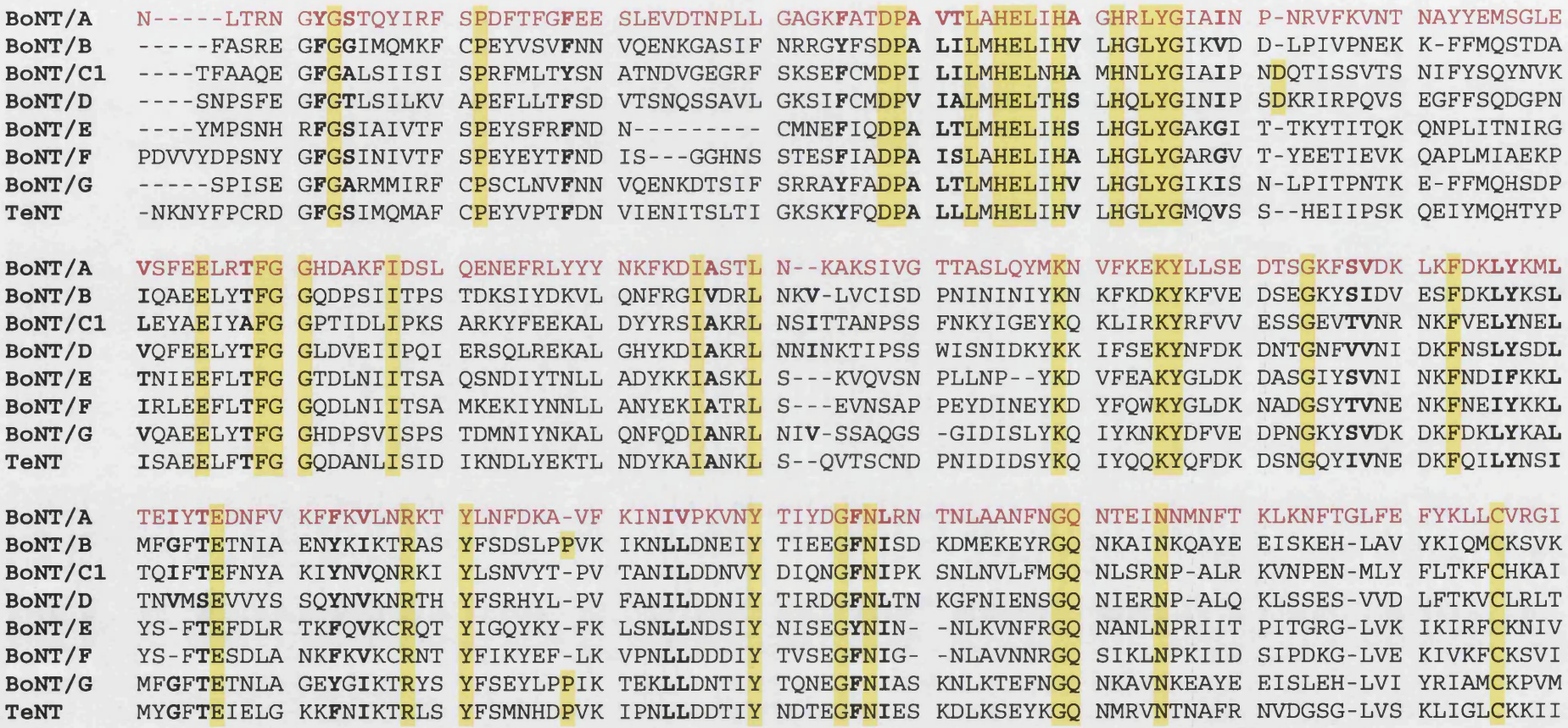
## Sequence Alignment of the Eight Clostridial Neurotoxins

The amino acid sequences of botulinum neurotoxins A-G and tetanus neurotoxin were obtained from the Entrez SeqSearch database, (www.ncbi.nim.nih.gov:80/entrez) and aligned using clustalw 1.8 (Thompson et al. 2003).

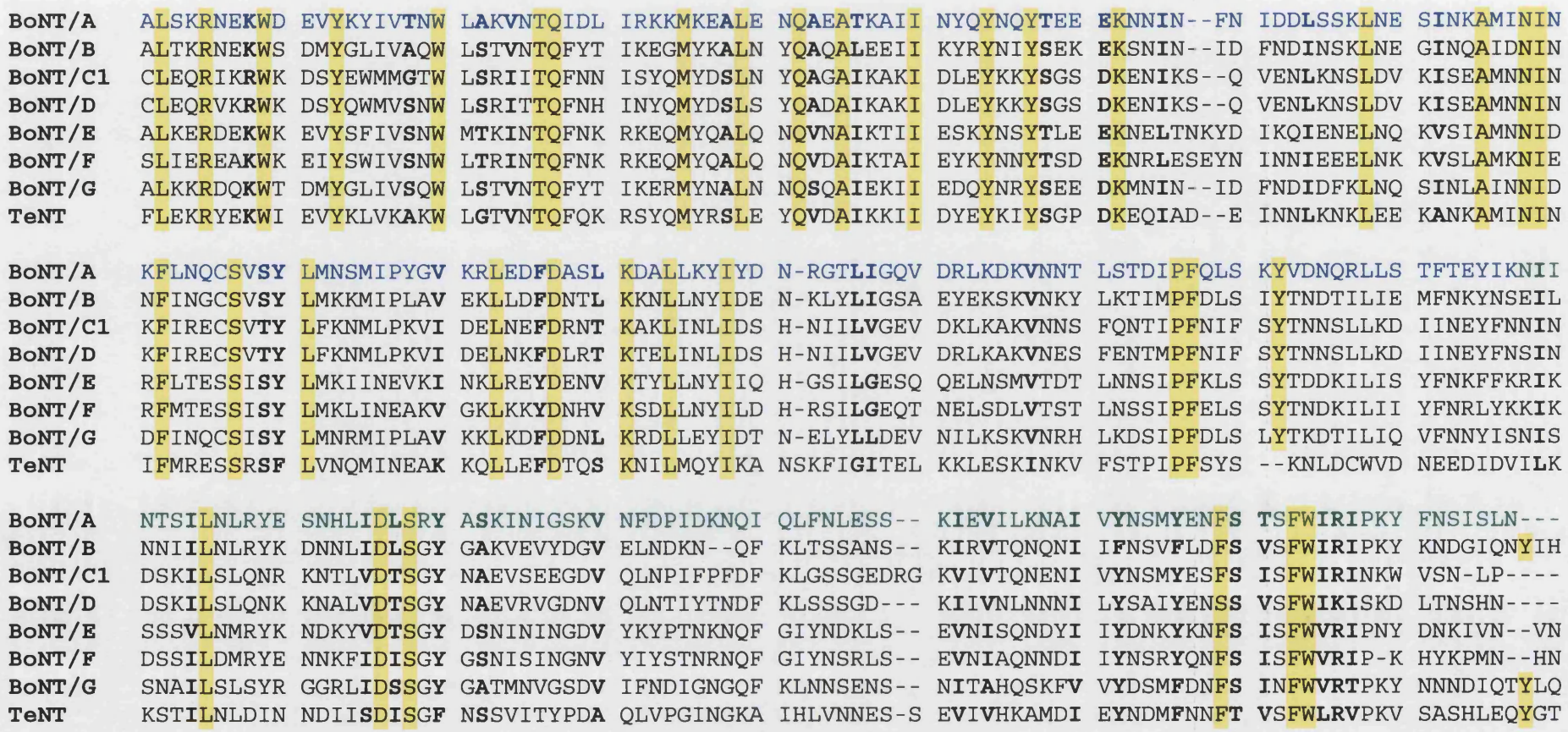
There is 30-40% sequence similarity among the clostridial neurotoxins. The three functional domains of BoNT/A have been coloured separately - the catalytic domain in red; the translocation domain in blue; and binding domain in green. Residues that are completely conserved among all the sequences are highlighted in yellow and conservatively substituted residues are indicated in bold type. The majority of conserved residues are located within the catalytic and translocation domains.

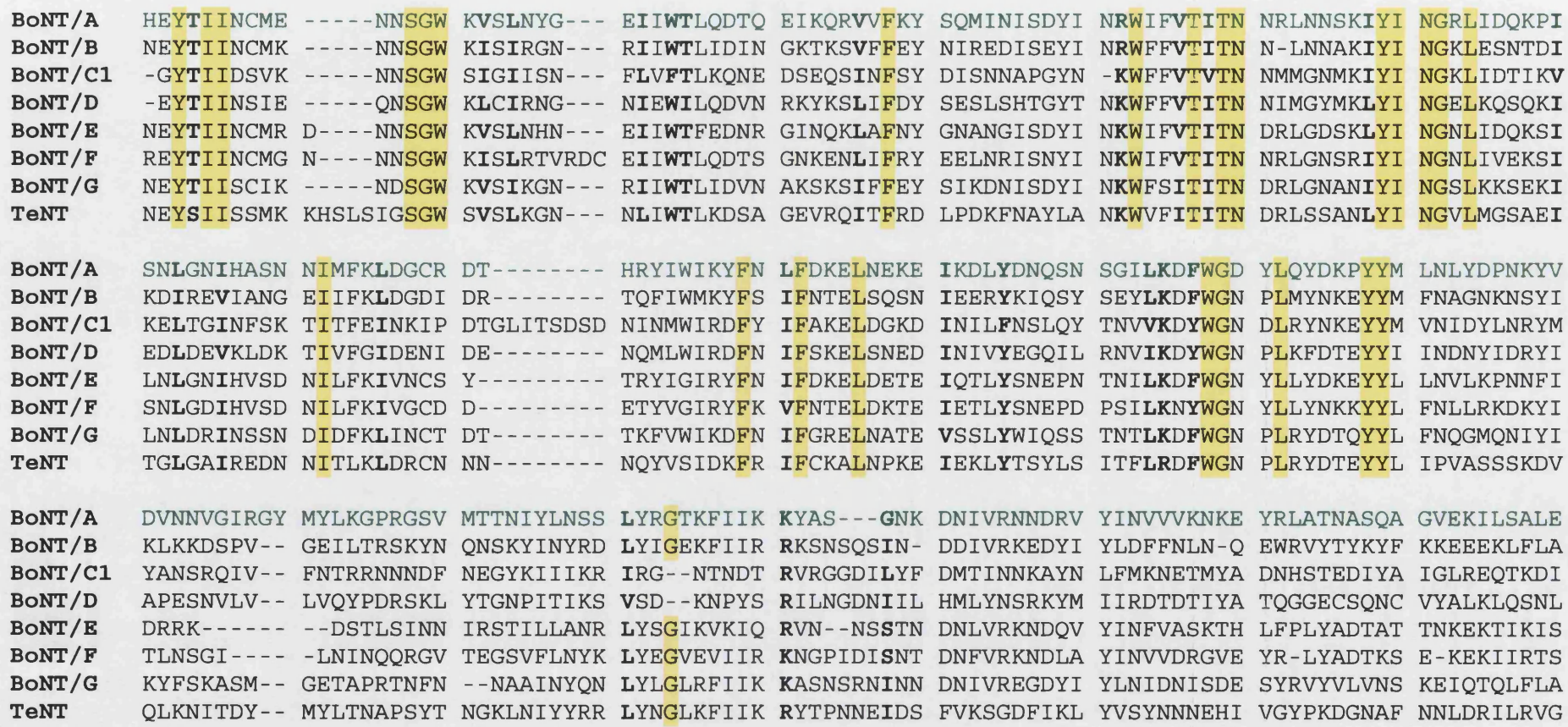
```
BoNT/A   -PFVNKQFNY KDPVNGVDIA YIKIPN-VGQ MQPVKAFKIH NKIWVIPERD TFTNPEEGDL NPPPEAKQVP VSYYDSTYLS TDNEKDNYLK
BoNT/B   -PVTINNFNY NDPIDNNNII MMEPPFARGT GRYYKAFKIT DRIWIIPERY TFGYKPEDFN KSSGIFNRDV CEYYDPDYLN TNDKKNIFLQ
BoNT/C1  -PITINNFNY SDPVDNKNIL YLDTHLNTLA NEPEKAFRIT GNIWVIPDRF SRNSNPNLNK PPRVTSPKSG --YYDPNYLS TDSDKDPFLK
BoNT/D   MTWPVKDFNY SDPVNDNDIL YLRIPQNKLI TTPVKAFMIT QNIWVIPERF SSDTNPSLSK PPRPTSKYQS --YYDPSYLS TDEQKDTFLK
BoNT/E   --PKINSFNY NDPVNDRTIL YIKPGG---C QEFYKSFNIM KNIWIIPERN VIGTTPQDFH PPTSLKNGDS -SYYDPNYLQ SDEEKDRFLK
BoNT/F   MPVAINSFNY NDPVNDDTIL YMQIPYEEKS KKYYKAFEIM RNVWIIPERN TIGTNPSDFD PPASLKNGSS -AYYDPNYLT TDAEKDRYLK
BoNT/G   -PVNIKXFNY NDPINNDDII MMEPFNDPGP GTYYKAFRII DRIWIVPERF TYGFQPDQFN ASTGVFSKDV YEYYDPTYLK TDAEKDKFLK
TeNT     -PITINNFRY SDPVNNDTII MMEPPYCKGL DIYYKAFKIT DRIWIVPERY EFGTKPEDFN PPSSLIEGAS -EYYDPNYLR TDSDKDRFLQ

BoNT/A   GVTKLFERIY STDLGRMLLT SIVRGIPFWG GSTIDTELKV IDTNCINVIQ PDGSYRSE-- ----ELNLVI IGPSADIIQF ECKSFGHEVL
BoNT/B   TMIKLFNRIK SKPLGEKLLE MIINGIPYLG DRRVPLEEFN TNIASVTVNK LISNPGEVER KKGIFANLII FGPGPVLNEN ETIDIGIQNH
BoNT/C1  EIIKLFKRIN SREIGEELIY RLSTDIPFPG NNNTPINTFD FDVDFNSVDV KTRQGNNWVK TGSINPSVII TGPRENIIDP ETSTFKLTNN
BoNT/D   GIIKLFKRIN ERDIGKKLIN YLVVGSPFMG DSSTPEDTFD FTRHTTNIAV EKFENGSWKV TNIITPSVLI FGPLPNILDY TASLTLQGQQ
BoNT/E   IVTKIFNRIN NNLSGGILLE ELSKANPYLG NDNTPDNQFH IG-DASAVEI KFSNGSQD-- --ILLPNVII MGAEPDLFET NSSNISLRNN
BoNT/F   TTIKLFKRIN SNPAGKVLLQ EISYAKPYLG NDHTPIDEFS PVTRTTSVNI KLSTNVES-- --SMLLNLLV LGAGPDIFES CCYPVRKLID
BoNT/G   TMIKLFNRIN SKPSGQRLLD MIVDAIPYLG NASTPPDKFA ANVANVSINK KIIQPGAEDQ IKGLMTNLII FGPGPVLSDN FTDSMIMNGH
TeNT     TMVKLFNRIK NNVAGEALLD KIINAIPYLG NSYSLLDKFD TNSNSVSFNL LEQDPSGATT KSAMLTNLII FGPGPVLNKN EVRGIVLRVD
```


BoNT/A N-----LTRN GYGSTQYIRF SPDFTFGFEE SLEVDTNPLL GAGKFATDPA VTLAHELIHA GHRLYGIAIN P-NRVFKVNT NAYYEMSGLE
BoNT/B -----FASRE GFGGIMQMKF CPEYVSVFNN VQENKGASIF NRRGYFSDPA LILMHELIHV LHGLYGIKVD D-LPIVPNEK K-FFMQSTDA
BoNT/C1 ----TFAAQE GFGALSIISI SPRFMLTYSN ATNDVGEGRF SKSEFCMDPI LILMHELNHA MHNLYGIAIP NDQTISSVTS NIFYSQYNVK
BoNT/D ----SNPSFE GFGTLSILKV APEFLLTFSD VTSNQSSAVL GKSIFCMDPV IALMHELTHS LHQLYGINIP SDKRIRPQVS EGFFSQDGPN
BoNT/E ----YMPSNH RFGSIAIVTF SPEYSFRFND N--------- CMNEFIQDPA LTLMHELIHS LHGLYGAKGI T-TKYTITQK QNPLITNIRG
BoNT/F PDVVYDPSNY GFGSINIVTF SPEYEYTFND IS---GGHNS STESFIADPA ISLAHELIHA LHGLYGARGV T-YEETIEVK QAPLMIAEKP
BoNT/G -----SPISE GFGARMMIRF CPSCLNVFNN VQENKDTSIF SRRAYFADPA LTLMHELIHV LHGLYGIKIS N-LPITPNTK E-FFMQHSDP
TeNT -NKNYFPCRD GFGSIMQMAF CPEYVPTFDN VIENITSLTI GKSKYFQDPA LLLMHELIHV LHGLYGMQVS S--HEIIPSK QEIYMQHTYP
BoNT/A VSFEELRTFG GHDAKFIDSL QENEFRLYYY NKFKDIASTL N--KAKSIVG TTASLQYMKN VFKEKYLLSE DTSGKFSVDK LKFDKLYKML
BoNT/B IQAEELYTFG GQDPSIITPS TDKSIYDKVL QNFRGIVDRL NKV-LVCISD PNININIYKN KFKDKYKFVE DSEGKYSIDV ESFDKLYKSL
BoNT/C1 LEYAEIYAFG GPTIDLIPKS ARKYFEEKAL DYYRSIAKRL NSITTANPSS FNKYIGEYKQ KLIRKYRFVV ESSGEVTVNR NKFVELYNEL
BoNT/D VQFEELYTFG GLDVEIIPQI ERSQLREKAL GHYKDIAKRL NNINKTIPSS WISNIDKYKK IFSEKYNFDK DNTGNFVVNI DKFNSLYSDL
BoNT/E TNIEEFLTFG GTDLNIITSA QSNDIYTNLL ADYKKIASKL S---KVQVSN PLLNP--YKD VFEAKYGLDK DASGIYSVNI NKFNDIFKKL
BoNT/F IRLEEFLTFG GQDLNIITSA MKEKIYNNLL ANYEKIATRL S---EVNSAP PEYDINEYKD YFQWKYGLDK NADGSYTVNE NKFNEIYKKL
BoNT/G VQAEELYTFG GHDPSVISPS TDMNIYNKAL QNFQDIANRL NIV-SSAQGS -GIDISLYKQ IYKNKYDFVE DPNGKYSVDK DKFDKLYKAL
TeNT ISAEELFTFG GQDANLISID IKNDLYEKTL NDYKAIANKL S--QVTSCND PNIDIDSYKQ IYQQKYQFDK DSNGQYIVNE DKFQILYNSI
BoNT/A TEIYTEDNFV KFFKVLNRKT YLNFDKA-VF KINIVPKVNY TIYDGFNLRN TNLAANFNGQ NTEINNMNFT KLKNFTGLFE FYKLLCVRGI
BoNT/B MFGFTETNIA ENYKIKTRAS YFSDSLPPVK IKNLLDNEIY TIEEGFNISD KDMEKEYRGQ NKAINKQAYE EISKEH-LAV YKIQMCKSVK
BoNT/C1 TQIFTEFNYA KIYNVQNRKI YLSNVYT-PV TANILDDNVY DIQNGFNIPK SNLNVLFMGQ NLSRNP-ALR KVNPEN-MLY FLTKFCHKAI
BoNT/D TNVMSEVVYS SQYNVKNRTH YFSRHYL-PV FANILDDNIY TIRDGFNLTN KGFNIENSGQ NIERNP-ALQ KLSSES-VVD LFTKVCLRLT
BoNT/E YS-FTEFDLR TKFQVKCRQT YIGQYKY-FK LSNLLNDSIY NISEGYNIN- -NLKVNFRGQ NANLNPRIIT PITGRG-LVK IKIRFCKNIV
BoNT/F YS-FTESDLA NKFKVKCRNT YFIKYEF-LK VPNLLDDDIY TVSEGFNIG- -NLAVNNRGQ SIKLNPKIID SIPDKG-LVE KIVKFCKSVI
BoNT/G MFGFTETNLA GEYGIKTRYS YFSEYLPPIK TEKLLDNTIY TQNEGFNIAS KNLKTEFNGQ NKAVNKEAYE EISLEH-LVI YRIAMCKPVM
TeNT MYGFTEIELG KKFNIKTRLS YFSMNHDPVK IPNLLDDTIY NDTEGFNIES KDLKSEYKGQ NMRVNTNAFR NVDGSG-LVS KLIGLCKKII


BoNT/A ITSKTKSLDK G----YNKAL NDLCIKVNNW DLFFSPSEDN FTNDLNKGEE ITSDTNIEAA EENISLDLIQ QYYLTFNFDN EPENISIENL
BoNT/B AP-------- ---------- -GICIDVDNE DLFFIADKNS FSDDLSKNER IEYNTQSNYI ENDFPINELI LDTDLISKIE LPSENTESLT
BoNT/C1 DGRSL----- -----YNKTL DCRELLVKNT DLPFIGDISD VKTDIFLRKD INEETEVIYY PDNVSVDQVI LS-KNTSEHG QLDLLYPSID
BoNT/D K--------- ------NSRD DSTCIKVKNN RLPYVADKDS ISQEIFENKI ITDETNVQNY SDKFSLDESI LDGQVPINPE IVDPLLPNVN
BoNT/E SVKG------ ------IRKS --ICIEINNG ELFFVASENS YNDDNINTPK EIDDTVTSNN NYENDLDQVI LNFNSESAPG LSDEKLNLTI
BoNT/F PRKG------ ------TKAP PRLCIRVNNS ELFFVASESS YNENDINTPK EIDDTTNLNN NYRNNLDEVI LDYNSQTIPQ ISNRTLNTLV
BoNT/G YKN------- -------TGK SEQCIIVNNE DLFFIANKDS FSKDLAKAET IAYNTQNNTI ENNFSIDQLI LDNDLSSGID LPNENTEPFT
TeNT PPTNIRENLY NRTASLTDLG GELCIKIKNE DLTFIAEKNS FSEEPFQDEI VSYNTKNKPL NFNYSLDKII VDYNLQSKIT LPNDRTTPVT
BoNT/A SSDIIGQLEL MPNIERFPNG KKYELDKYTM FHYLRAQEFE HGKSRIALTN SVNEALLNPS RVYTFFSSDY VKKVNKATEA AMFLGWVEQL
BoNT/B DFN-V--DVP VYEKQ--PAI KKIFTDENTI FQYLYSQTFP LDIRDISLTS SFDDALLFSN KVYSFFSMDY IKTANKVVEA GLFAGWVKQI
BoNT/C1 SES----EIL PGENQVFYDN RTQNVDYLNS YYYLESQKLS DNVEDFTFTR SIEEALDNSA KVYTYFPTLA NKVN-AGVQG GLFLMWANDV
BoNT/D MEP----LNL PGEEIVFYDD ITKYVDYLNS YYYLESQKLS NNVENITLTT SVEEALGYSN KIYTFLPSLA EKVN-KGVQA GLFLNWANEV
BoNT/E QND-A--YIP KYDSNGTSDI EQHDVNELNV FFYLDAQKVP EGENNVNLTS SIDTALLEQP KIYTFFSSEF INNVNKPVQA ALFVSWIQQV
BoNT/F QDN-S--YVP RYDSNGTSEI EEYDVVDFNV FFYLHAQKVP EGETNISLTS SIDTALLEES -KDIFFSSEF IDTINKPVNA ALFIDWISKV
BoNT/G NFDDI--DIP VYIKQ--SAL KKIFVDGDSL FEYLHAQTFP SNIENLQLTN SLNDALRNNN KVYTFFSTNL VEKANTVVGA SLFVNWVKGV
TeNT KGIP---YAP EYKSNAASTI EIHNIDDNTI YQYLYAQKSP TTLQRITMTN SVDDALINST KIYSYFPSVI SKVN-QGAQG ILFLQWVRDI
BoNT/A VYDFTDETSE VSTTDKIADI TIIIPYIGPA LNIGNMLYKD DFVGALIFSG AVILLEFIPE IAIPVLGTFA LVSYIAN--- KVLTVQTIDN
BoNT/B VNDFVIEANK SNTMDKIADI SLIVPYIGLA LNVGNETAKG NFENAFEIAG ASILLEFIPE LLIPVVGAFL LESYID---N KNKIIKTIDN
BoNT/C1 VEDFTTNILR KDTLDKISDV SAIIPYIGPA LNISNSVRRG NFTEAFAVTG VTILLEAFPE FTIPALGAFV IYSKVQ---E RNEIIKTIDN
BoNT/D VEDFTTNIMK KDTLDKISDV SVIIPYIGPA LNIGNSALRG NFNQAFATAG VAFLLEGFPE FTIPALGVFT FYSSIQ---E REKIIKTIEN
BoNT/E LVDFTTEANQ KSTVDKIADI SIVVPYIGLA LNIGNEAQKG NFKDALELLG AGILLEFEPE LLIPTILVFT IKSFLGSSDN KNKVIKAINN
BoNT/F IRDFTTEATQ KSTVDKIADI SLIVPYVGLA LNIIIEAEKG NFEEAFELLG VGILLEFVPE LTIPVILVFT IKSYIDSYEN KNKAIKAINN
BoNT/G IDDFTSESTQ KSTIDKVSDV SIIIPYIGPA LNVGNETAKE NFKNAFEIGG AAILMEFIPE LIVPIVGFFT LESYVG---N KGHIIMTISN
TeNT IDDFTNESSQ KTTIDKISDV STIVPYIGPA LNIVKQGYEG NFIGALETTG VVLLLEYIPE ITLPVIAALS IAESST---Q KEKIIKTIDN


BoNT/A ALSKRNEKWD EVYKYIVTNW LAKVNTQIDL IRKKMKEALE NQAEATKAII NYQYNQYTEE EKNNIN--FN IDDLSSKLNE SINKAMININ
BoNT/B ALTKRNEKWS DMYGLIVAQW LSTVNTQFYT IKEGMYKALN YQAQALEEII KYRYNIYSEK EKSNIN--ID FNDINSKLNE GINQAIDNIN
BoNT/C1 CLEQRIKRWK DSYEWMMGTW LSRIITQFNN ISYQMYDSLN YQAGAIKAKI DLEYKKYSGS DKENIKS--Q VENLKNSLDV KISEAMNNIN
BoNT/D CLEQRVKRWK DSYQWMVSNW LSRITTQFNH INYQMYDSLS YQADAIKAKI DLEYKKYSGS DKENIKS--Q VENLKNSLDV KISEAMNNIN
BoNT/E ALKERDEKWK EVYSFIVSNW MTKINTQFNK RKEQMYQALQ NQVNAIKTII ESKYNSYTLE EKNELTNKYD IKQIENELNQ KVSIAMNNID
BoNT/F SLIEREAKWK EIYSWIVSNW LTRINTQFNK RKEQMYQALQ NQVDAIKTAI EYKYNNYTSD EKNRLESEYN INNIEEELNK KVSLAMKNIE
BoNT/G ALKKRDQKWT DMYGLIVSQW LSTVNTQFYT IKERMYNALN NQSQAIEKII EDQYNRYSEE DKMNIN--ID FNDIDFKLNQ SINLAINNID
TeNT FLEKRYEKWI EVYKLVKAKW LGTVNTQFQK RSYQMYRSLE YQVDAIKKII DYEYKIYSGP DKEQIAD--E INNLKNKLEE KANKAMININ
BoNT/A KFLNQCSVSY LMNSMIPYGV KRLEDFDASL KDALLKYIYD N-RGTLIGQV DRLKDKVNNT LSTDIPFQLS KYVDNQRLLS TFTEYIKNII
BoNT/B NFINGCSVSY LMKKMIPLAV EKLLDFDNTL KKNLLNYIDE N-KLYLIGSA EYEKSKVNKY LKTIMPFDLS IYTNDTILIE MFNKYNSEIL
BoNT/C1 KFIRECSVTY LFKNMLPKVI DELNEFDRNT KAKLINLIDS H-NIILVGEV DKLKAKVNNS FQNTIPFNIF SYTNNSLLKD IINEYFNNIN
BoNT/D KFIRECSVTY LFKNMLPKVI DELNKFDLRT KTELINLIDS H-NIILVGEV DRLKAKVNES FENTMPFNIF SYTNNSLLKD IINEYFNSIN
BoNT/E RFLTESSISY LMKIINEVKI NKLREYDENV KTYLLNYIIQ H-GSILGESQ QELNSMVTDT LNNSIPFKLS SYTDDKILIS YFNKFFKRIK
BoNT/F RFMTESSISY LMKLINEAKV GKLKKYDNHV KSDLLNYILD H-RSILGEQT NELSDLVTST LNSSIPFELS SYTNDKILII YFNRLYKKIK
BoNT/G DFINQCSISY LMNRMIPLAV KKLKDFDDNL KRDLLEYIDT N-ELYLLDEV NILKSKVNRH LKDSIPFDLS LYTKDTILIQ VFNNYISNIS
TeNT IFMRESSRSF LVNQMINEAK KQLLEFDTQS KNILMQYIKA NSKFIGITEL KKLESKINKV FSTPIPFSYS --KNLDCWVD NEEDIDVILK
BoNT/A NTSILNLRYE SNHLIDLSRY ASKINIGSKV NFDPIDKNQI QLFNLESS-- KIEVILKNAI VYNSMYENFS TSFWIRIPKY FNSISLN---
BoNT/B NNIILNLRYK DNNLIDLSGY GAKVEVYDGV ELNDKN--QF KLTSSANS-- KIRVTQNQNI IFNSVFLDFS VSFWIRIPKY KNDGIQNYIH
BoNT/C1 DSKILSLQNR KNTLVDTSGY NAEVSEEGDV QLNPIFPFDF KLGSSGEDRG KVIVTQNENI VYNSMYESFS ISFWIRINKW VSN-LP----
BoNT/D DSKILSLQNK KNALVDTSGY NAEVRVGDNV QLNTIYTNDF KLSSSGD--- KIIVNLNNNI LYSAIYENSS VSFWIKISKD LTNSHN----
BoNT/E SSSVLNMRYK NDKYVDTSGY DSNININGDV YKYPTNKNQF GIYNDKLS-- EVNISQNDYI IYDNKYKNFS ISFWVRIPNY DNKIVN--VN
BoNT/F DSSILDMRYE NNKFIDISGY GSNISINGNV YIYSTNRNQF GIYNSRLS-- EVNIAQNNDI IYNSRYQNFS ISFWVRIP-K HYKPMN--HN
BoNT/G SNAILSLSYR GGRLIDSSGY GATMNVGSDV IFNDIGNGQF KLNNSENS-- NITAHQSKFV VYDSMFDNFS INFWVRTPKY NNNDIQTYLQ
TeNT KSTILNLDIN NDIISDISGF NSSVITYPDA QLVPGINGKA IHLVNNES-S EVIVHKAMDI EYNDMFNNFT VSFWLRVPKV SASHLEQYGT


BoNT/A HEYTIINCME -----NNSGW KVSLNYG--- EIIWTLQDTQ EIKQRVVFKY SQMINISDYI NRWIFVTITN NRLNNSKIYI NGRLIDQKPI
BoNT/B NEYTIINCMK -----NNSGW KISIRGN--- RIIWTLIDIN GKTKSVFFEY NIREDISEYI NRWFFVTITN N-LNNAKIYI NGKLESNTDI
BoNT/C1 -GYTIIDSVK -----NNSGW SIGIISN--- FLVFTLKQNE DSEQSINFSY DISNNAPGYN -KWFFVTVTN NMMGNMKIYI NGKLIDTIKV
BoNT/D -EYTIINSIE -----QNSGW KLCIRNG--- NIEWILQDVN RKYKSLIFDY SESLSHTGYT NKWFFVTITN NIMGYMKLYI NGELKQSQKI
BoNT/E NEYTIINCMR D----NNSGW KVSLNHN--- EIIWTFEDNR GINQKLAFNY GNANGISDYI NKWIFVTITN DRLGDSKLYI NGNLIDQKSI
BoNT/F REYTIINCMG N----NNSGW KISLRTVRDC EIIWTLQDTS GNKENLIFRY EELNRISNYI NKWIFVTITN NRLGNSRIYI NGNLIVEKSI
BoNT/G NEYTIISCIK -----NDSGW KVSIKGN--- RIIWTLIDVN AKSKSIFFEY SIKDNISDYI NKWFSITITN DRLGNANIYI NGSLKKSEKI
TeNT NEYSIISSMK KHSLSIGSGW SVSLKGN--- NLIWTLKDSA GEVRQITFRD LPDKFNAYLA NKWVFITITN DRLSSANLYI NGVLMGSAEI
BoNT/A SNLGNIHASN NIMFKLDGCR DT-------- HRYIWIKYFN LFDKELNEKE IKDLYDNQSN SGILKDFWGD YLQYDKPYYM LNLYDPNKYV
BoNT/B KDIREVIANG EIIFKLDGDI DR-------- TQFIWMKYFS IFNTELSQSN IEERYKIQSY SEYLKDFWGN PLMYNKEYYM FNAGNKNSYI
BoNT/C1 KELTGINFSK TITFEINKIP DTGLITSDSD NINMWIRDFY IFAKELDGKD INILFNSLQY TNVVKDYWGN DLRYNKEYYM VNIDYLNRYM
BoNT/D EDLDEVKLDK TIVFGIDENI DE-------- NQMLWIRDFN IFSKELSNED INIVYEGQIL RNVIKDYWGN PLKFDTEYYI INDNYIDRYI
BoNT/E LNLGNIHVSD NILFKIVNCS Y--------- TRYIGIRYFN IFDKELDETE IQTLYSNEPN TNILKDFWGN YLLYDKEYYL LNVLKPNNFI
BoNT/F SNLGDIHVSD NILFKIVGCD D--------- ETYVGIRYFK VFNTELDKTE IETLYSNEPD PSILKNYWGN YLLYNKKYYL FNLLRKDKYI
BoNT/G LNLDRINSSN DIDFKLINCT DT-------- TKFVWIKDFN IFGRELNATE VSSLYWIQSS TNTLKDFWGN PLRYDTQYYL FNQGMQNIYI
TeNT TGLGAIREDN NITLKLDRCN NN-------- NQYVSIDKFR IFCKALNPKE IEKLYTSYLS ITFLRDFWGN PLRYDTEYYL IPVASSSKDV
BoNT/A DVNNVGIRGY MYLKGPRGSV MTTNIYLNSS LYRGTKFIIK KYAS---GNK DNIVRNNDRV YINVVVKNKE YRLATNASQA GVEKILSALE
BoNT/B KLKKDSPV-- GEILTRSKYN QNSKYINYRD LYIGEKFIIR RKSNSQSIN- DDIVRKEDYI YLDFFNLN-Q EWRVYTYKYF KKEEEKLFLA
BoNT/C1 YANSRQIV-- FNTRRNNNDF NEGYKIIIKR IRG--NTNDT RVRGGDILYF DMTINNKAYN LFMKNETMYA DNHSTEDIYA IGLREQTKDI
BoNT/D APESNVLV-- LVQYPDRSKL YTGNPITIKS VSD--KNPYS RILNGDNIIL HMLYNSRKYM IIRDTDTIYA TQGGECSQNC VYALKLQSNL
BoNT/E DRRK------ --DSTLSINN IRSTILLANR LYSGIKVKIQ RVN--NSSTN DNLVRKNDQV YINFVASKTH LFPLYADTAT TNKEKTIKIS
BoNT/F TLNSGI---- -LNINQQRGV TEGSVFLNYK LYEGVEVIIR KNGPIDISNT DNFVRKNDLA YINVVDRGVE YR-LYADTKS E-KEKIIRTS
BoNT/G KYFSKASM-- GETAPRTNFN --NAAINYQN LYLGLRFIIK KASNSRNINN DNIVREGDYI YLNIDNISDE SYRVYVLVNS KEIQTQLFLA
TeNT QLKNITDY-- MYLTNAPSYT NGKLNIYYRR LYNGLKFIIK RYTPNNEIDS FVKSGDFIKL YVSYNNNEHI VGYPKDGNAF NNLDRILRVG

```
BoNT/A   IPDVGNLSQV VVMKSKNDQG ITNKCKMNLQ DNNGNDIGFI G--------- -FHQFNNIAK LVASNWYNRQ IERSSR--TL GCSWEFIPVD
BoNT/B   PISDSDEFYN TIQIKEYDEQ PTYSCQLLFK KDEESTDEIG LIGIHRFYES GIVFEEYKDY FCISKWYLKE VKRKPYNLKL GCNWQFIPKD
BoNT/C1  NDN--IIFQI QPMNNTYYYA SQIFKSNFNG ENISGIC--- ---------- --SIGTYRFR LGGDWYRHNY LVPTVKQGNY ASLLESTSTH
BoNT/D   GNYGIGIFSI KNIVSKNKYC SQIFSS-FRE NTMLLAD--- ---------- --IYKPWRFS FKNAYTPVAV TNYETKLLST SSFWKFISRD
BoNT/E   SSGNR--FNQ VVVMNSVG-N CTMNFKNNNG NNIG------ ---------- --LLGFKADT VVASTWYYTH MRDHTN--SN GCFWNFISEE
BoNT/F   NLNDS--LGQ IIVMDSIGNN CTMNFQNNNG SNIG------ ---------- --LLGFHSNN LVASSWYYNN IRRNTS--SN GCFWSSISKE
BoNT/G   PINDDPTFYD VLQIKKYYEK TTYNCQILCE KDTKTFGLFG IG--KFVKDY GYVWDTYDNY FCISQWYLRR ISENINKLRL GCNWQFIPVD
TeNT     YNAPGIPLYK KMEAVKLRDL KTYSVQLKLY DDKNASLGLV G-----THNG QIGNDPNRDI LIASNWYFNH LKDKILG--- -CDWYFVPTD

BoNT/A   DGWGERPL
BoNT/B   EGWTE---
BoNT/C1  WGFVPVSE
BoNT/D   PGWVE---
BoNT/E   HGWQEK—
BoNT/F   NGWKE---
BoNT/G   EGWTE---
TeNT     EGWTND--
```

**Flowchart Outlining the Steps Involved in the Recombinant Expression and Purification of $LH_N/A$**

(as described in Chaddock et al, 2002)

The pGEX-4T-2 expression vector was modified to incorporate strong *rrn*B transcriptional terminators downstream of the cloning site and the strong *trc* promoter.

↓

The expression cassette was constructed using three sources of DNA:
Bases 1-913 – synthetic DNA exhibiting an *E.coli* codon bias;
Bases 914-1138 and 1977-2616 – derived from PCR amplification from clostridial DNA;
Bases 1139-1976 – obtained by direct cloning of chromosomal DNA.

↓

pGSTLH$_N$/A was transformed into *E.coli* TG1 cells, which were grown in Terrific broth at 30°C to an $OD_{600}$ of 8.0. Expression of *rec*GSTLH$_N$/A was induced by the addition of 0.5mM IPTG and a reduction in temperature to 25°C.

↓

After 4 hours, the cells were harvested by centrifugation. Purification of *rec*LH$_N$/A from the cell paste utilised ion-exchange and affinity FPLC separation techniques.

↓

The initial capture step consisted of a glutathione-Sepharose 4B column immediately preceded by a Q-Sepharose-FF column connected in tandem. The glutathione-Sepharose 4B column was eluted with 10mM reduced glutathione and the resulting fractions were pooled and dialysed against 25mM Bis-Tris, pH 6.0.

↓

The dialysed material was treated with 0.1mg trypsin/mg protein for 30 mins at 37°C to remove the GST expression tag and cleave the LC-H$_N$ junction (thus activating LH$_N$/A).

↓

Trypsin-treated material was purified using a Q-Sepharose 4B column in tandem with a glutathione-Sepharose 4B column. The material was passed through the glutathione-Sepharose column first to remove the GST tag and uncleaved, GST-tagged protein. *rec*LH$_N$/A-containing fractions were pooled, concentrated and stored at -20°C.

**Conditions Reported for the Crystallization of a Variety of Clostridial Neurotoxins and Clostridial Neurotoxin Fragments**

| Protein | Mother Liquor | Reference |
|---|---|---|
| BoNT/A | 1.5M sodium formate | Stevens et al. 1991 |
| BoNT/A | 10% PEG 4000<br>0.2M magnesium chloride | Stevens et al. 1991 |
| BoNT/A | 0.8M sodium potassium tartrate<br>0.1M Hepes, pH 8.5 | Stevens et al. 1991 |
| BoNT/A | 0.15M magnesium acetate<br>11% PEG 4000<br>0.1M Tris, pH 7.0 | Lacy et al. 1998 |
| BoNT/A | 30% PEG 600<br>0.1M Tris, pH 7.0 | Lacy and Stevens 1999 |
| BoNT/A-$H_N$ | 0.1M Tris, pH 8.5<br>0.1M magnesium formate<br>1mM sodium azide | Lacy and Stevens 1997 |
| BoNT/A-LC minimal essential domain | 25% PEG 4000<br>0.1M sodium acetate, pH 4.6<br>3% xylitol | Kadkhodayan et al. 2000 |
| BoNT/B | 5mM citric acid, pH 5.6<br>10mM sodium phosphate | Flicker et al. 1999 |
| BoNT/B | PEG 6000<br>0.1M MES pH 6.0 | Swaminathan and Eswaramoorthy 2000a |
| BoNT/B | 10% PEG 4000<br>0.1M MES, pH 6.0 | Swaminathan and Eswaramoorthy 2000b |
| BoNT/B-synaptobrevin II | 0.1M Hepes, pH 7.0-7.5<br>15-25% PEG 4000 | Hanson and Stevens 2000 |
| TeNT-$H_C$ | 0.2M ammonium sulphate<br>40% PEG 4000<br>1% MPD | Emsley et al. 2000 |
| TeNT-$H_C$ | 20% PEG 4000<br>0.2M imidazole-malate, pH 7.0 | Fotinou et al. 2001 |
| TeNT-$H_C$ | 20% PEG 8000<br>0.2M sodium potassium phosphate<br>0.1M Tris, pH 8.5 | Fotinou et al. 2001 |

These conditions, reported for several clostridial neurotoxins and clostridial neurotoxin fragments, were used in crystallization experiments on $LH_N/A$ and $LH_N/A(H^{227}Y)$.

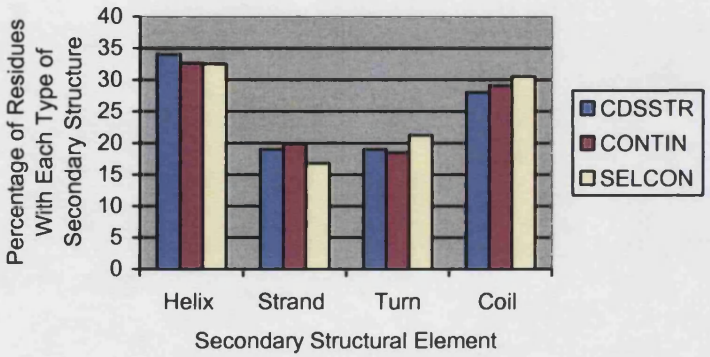
**Conditions Reported for the Crystallization of Zinc Proteases**

| Mother Liquor | Protein | Reference |
|---|---|---|
| 0.1M Hepes, pH 7.0<br>0.1M potassium thiocyanate<br>4.5M sodium chloride | Aminopeptidase | Chevrier et al (1994) Structre, **2**, 283-291 |
| 0.1-3.0M sodium chloride<br>10mM imidazole pH 7.0 | Aspzincin | Hori et al (2001) Acta Crystallogr, **D57**, 361-368 |
| 22% PEG 4000<br>5-10% glycerol<br>5-10mM calcium chloride | DNA polymerase III | Barnes et al (1998) Biochemistry, **37**, 15254-15260 |
| 0.1M imidazole, pH 6.8<br>10mM calcium chloride<br>2.4M ammonium sulphate | Atrolysin C | Zhang et al (1994) PNAS, **91**, 8447-8451 |
| 5-12% PEG 4000<br>10% glycerol<br>20mM zinc sulphate | DNA polymerase III | Barnes et al (1998) Biochemistry, **37**, 15254-15260 |
| 20% PEG 8000<br>0.1M zinc acetate<br>0.1M Pipes, pH 6.1 | Nerve growth factor | Holland et al (1994) J Mol Biol, **239**, 385-400 |
| 30% PEG 8000<br>0.1M ammonium sulphate<br>0.1M sodium citrate, pH 6.3 | 50kDa metalloprotease | Baumann (1994) J Mol Biol, **242**, 244-251 |
| 20-24% PEG 8000<br>2.5% isopropanol<br>10mM calcium chloride<br>0.1M Tris-HCl, pH 7.5-8.5 | Stromelysin catalytic domain | Chen et al (1999) J Mol Biol, **293**, 545-557 |
| 0.2M ammonium sulphate<br>0.1M sodium cacodylate, pH 6.5<br>15-30% PEG 8000 | Collagenase catalytic domain | Lovejoy et al (1994) Science, **263**, 375-377 |
| 0.1M Hepes, pH 7.5 or 8.2<br>0.3M sodium chloride<br>22% PEG 5000 MME | Human thrombin complexes | Bradley et al (2001) J Mol Biol, **301**, 1451-1486 |
| 0.1M sodium acetate, pH 4.6<br>2.0M sodium formate | Peptide deformylase | Chan et al (1997) Biochemistry, **36**, 13904-13909 |
| 50mM MES, pH 5.5<br>0.5M ammonium sulphate<br>1% trifluoroethanol | Stromelysin-3 catalytic domain | Gall et al (2001) J Mol Biol, **307**, 577-586 |
| 1.3M ammonium sulphate<br>0.1M Tris-HCl, pH 7.5 | 6-pyruvol tetrahydropterin synthase | Nar et al (1994) EMBO J, **13**, 1255-1262 |
| 2.0M sodium formate<br>1mM calcium chloride | Collagenase 19kDa catalytic domain | Lovejoy et al (1994) Biochemistry, **33**, 8207-8217 |

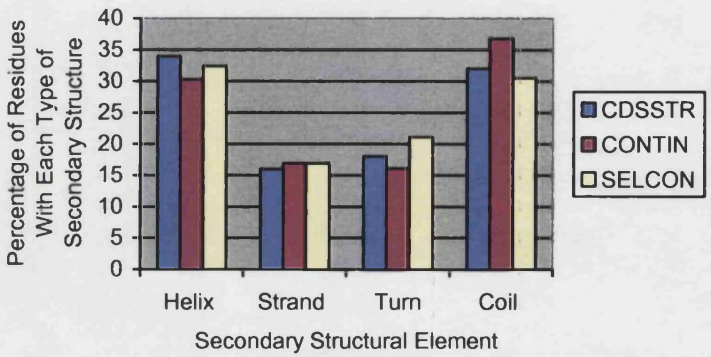
| | | |
|---|---|---|
| 50mM zinc chloride<br>0.1M Tris, pH 8 | | |
| 20% PEG 8000<br>25mM sodium citrate, pH 5.6<br>0.1M zinc acetate | β-lactamase | Carfi et al (1998) Acta Crystallogr, **D54**, 313-323 |
| 20% PEG 400<br>10% isopropanol | Glucose dehydrogenase | John. et al (1994) Structure, **2**, 385-393 |
| 20% PEG 4000<br>20% isopropanol<br>0.1M sodium citrate, pH 5.4 | Catalytic domain of TNFα-converting enzyme | Maskos et al (1998) PNAS, **95**, 3408-3412 |
| 20% PEG 4000<br>20% isopropanol<br>0.1M sodium citrate, pH 6.5 | Human uPA complexes | Bradley et al (2001) J Mol Biol, **301**, 1451-1486 |
| 20% PEG 4000<br>Tris-HCl, pH 7.5 | Acutolysin | Zhu et al, (1999) Acta Crystallogr, **D55**, 1834 |
| 20% PEG 6000<br>0.5M sodium chloride<br>0.1M calcium chloride<br>0.2M MES, ph 6.0<br>0.02% sodium azide | Collagenase catalytic domain | Bode et al (1994) EMBO J, **13**, 1263-1269 |

In addition, two drops were set up with mother liquor comprising 50mM sodium acetate, pH 4.6, 10mM zinc sulphate and 15% PEG 4000, as this had previously been used by another member of the lab to crystallize a zinc-containing protein.

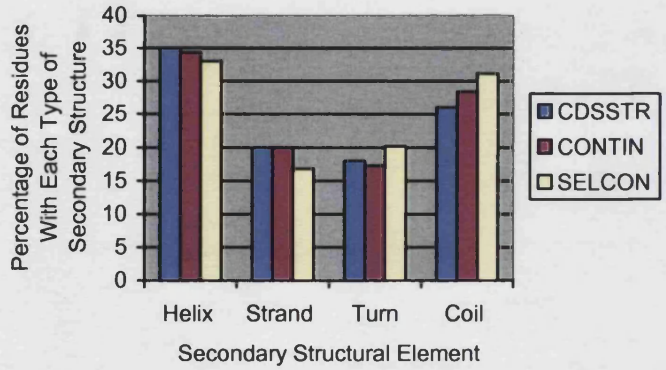
Secondary Structural Composition of LHN/A at pH 7.0
Analysed Using Protein Reference Set 3



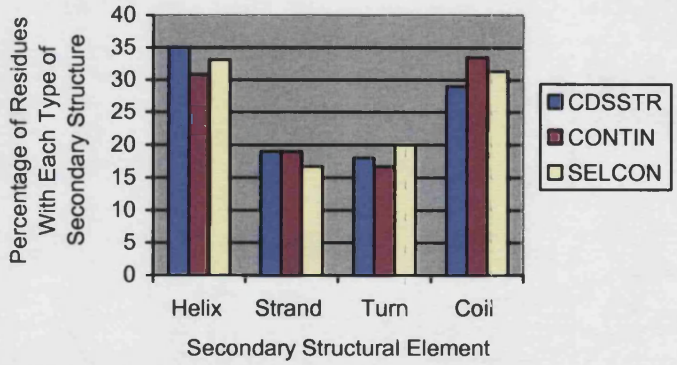

Secondary Structural Composition of LHN/A at pH 7.0
Analysed Using Protein Reference Set 6




Secondary Structural Composition of LHN/A(H227Y) at
pH 7.0 Analysed Using Protein Reference Set 3




Secondary Structural Composition of LHN/A(H227Y) at
pH 7.0 Analysed Using Protein Reference Set 6



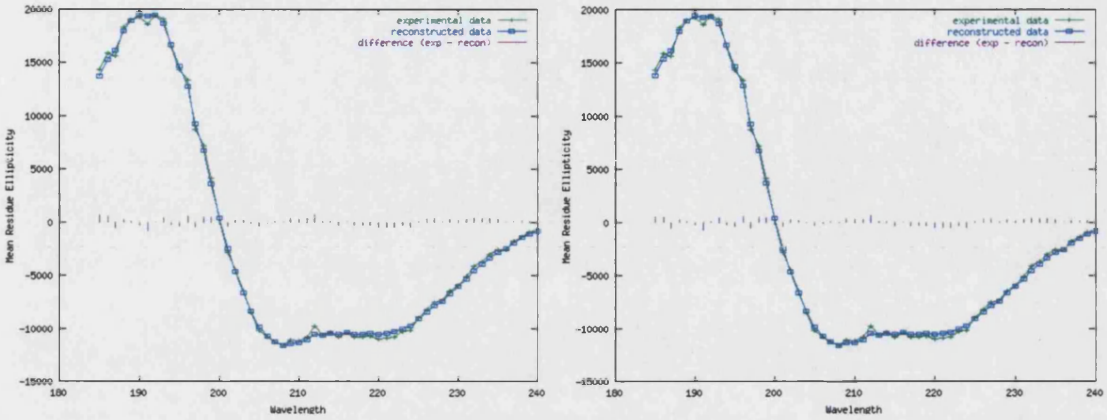
**Secondary Structure Compositions of $LH_N/A$ and $LH_N/A(H^{227}Y)$ at pH 7.0**

| Fragment | Analysis Program | Reference Set | α-Helix (%) | β-Strand (%) | β-Turn (%) | Random Coil (%) |
|---|---|---|---|---|---|---|
| **$LH_N/A$** | CDSSTR | 3 | 34 | 19 | 19 | 28 |
| | CONTINLL | 3 | 32.6 | 19.9 | 18.5 | 29.1 |
| | SELCON3 | 3 | 32.5 | 16.8 | 21.2 | 30.5 |
| | CDSSTR | 6 | 34 | 16 | 18 | 32 |
| | CONTINLL | 6 | 30.3 | 16.9 | 16.1 | 36.8 |
| | SELCON3 | 6 | 32.4 | 16.9 | 21.1 | 30.5 |
| **$LH_N/A(H^{227}Y)$** | CDSSTR | 3 | 35 | 20 | 18 | 26 |
| | CONTINLL | 3 | 34.3 | 20 | 17.3 | 28.4 |
| | SELCON3 | 3 | 33 | 16.8 | 20.2 | 31.1 |
| | CDSSTR | 6 | 35 | 19 | 18 | 29 |
| | CONTINLL | 6 | 30.8 | 19 | 16.7 | 33.5 |
| | SELCON3 | 6 | 33.1 | 16.7 | 20 | 31.3 |

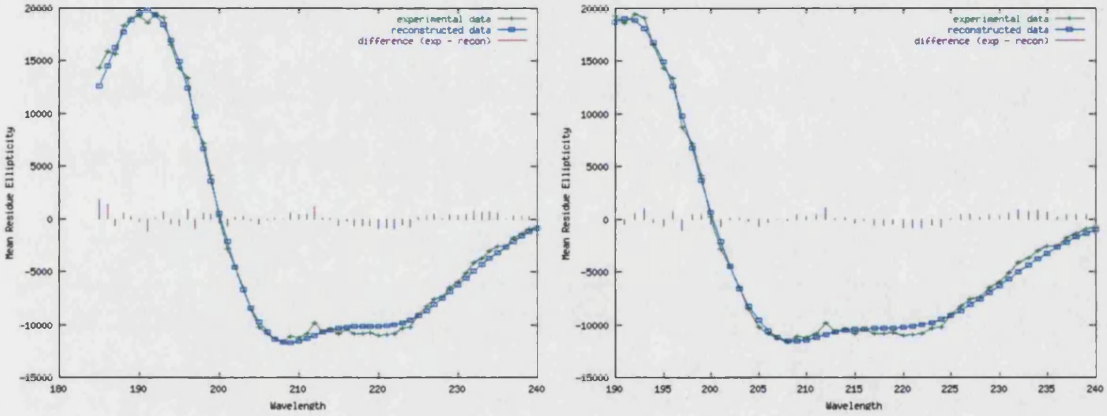
The percentage of residues involved in each of the four secondary structural elements (α-helix, β-strand, β-turn and random coil), as analysed by the programs CDSSTR, CONTINLL and SELCON3 with protein reference sets 3 and 6, are presented.

**Analysis of Far-UV CD Spectra Recorded for $LH_N/A$**

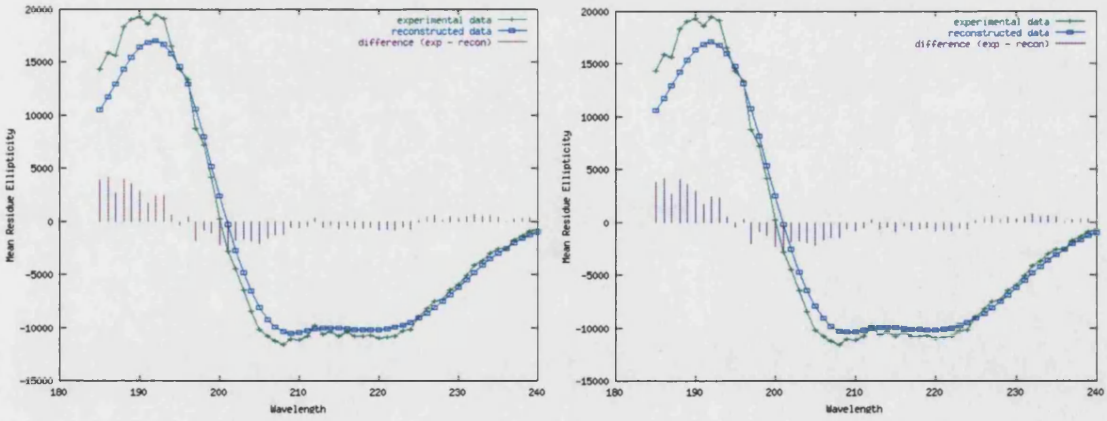
CDSSTR




CONTINLL



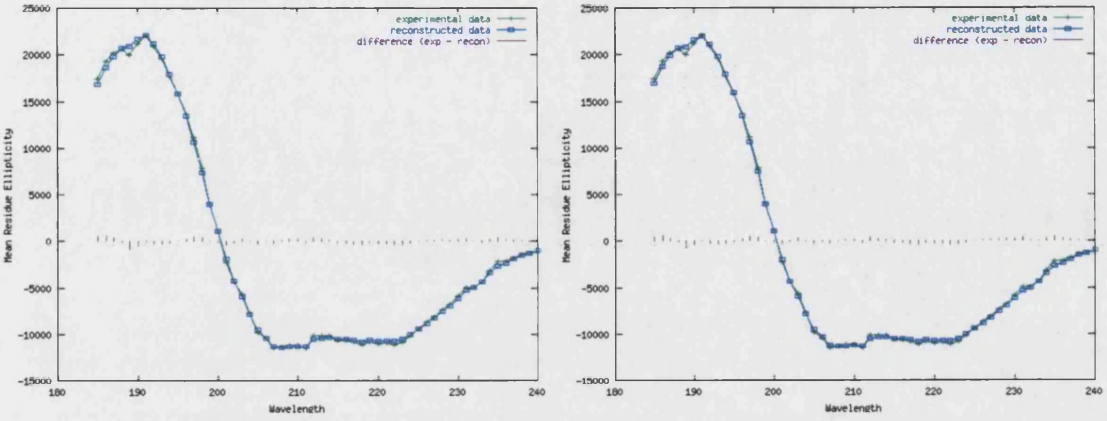

SELCON3



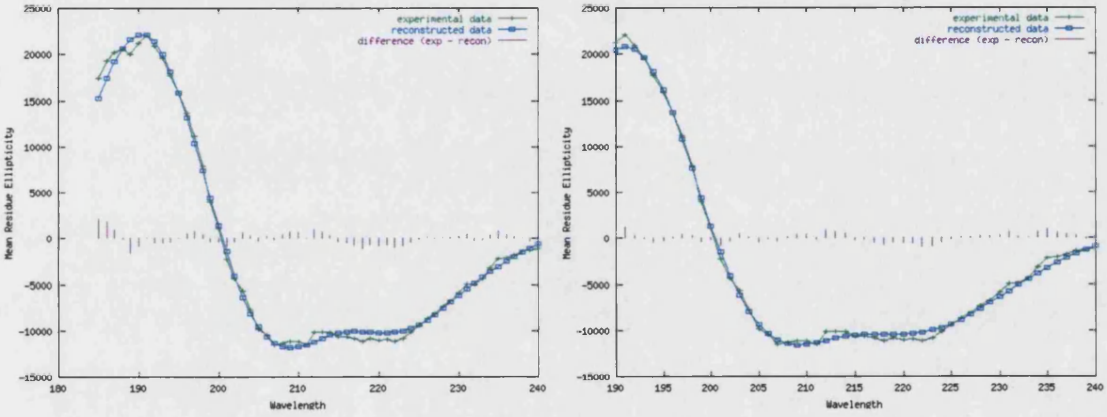

Graphical analyses of the CD spectra recorded for $LH_N/A$ using CDSSTR, CONTINLL and SELCON3. The panels on the left represent spectra analysed using protein reference set 3 and those on the right with reference set 6..

## Analysis of Far-UV CD Spectra Recorded for $LH_N/A(H^{227}Y)$

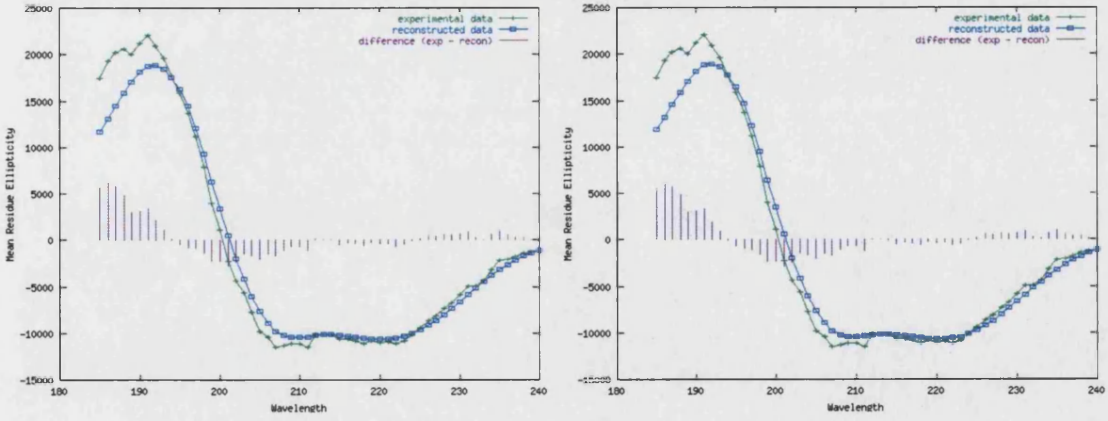
CDSSTR




CONTINLL

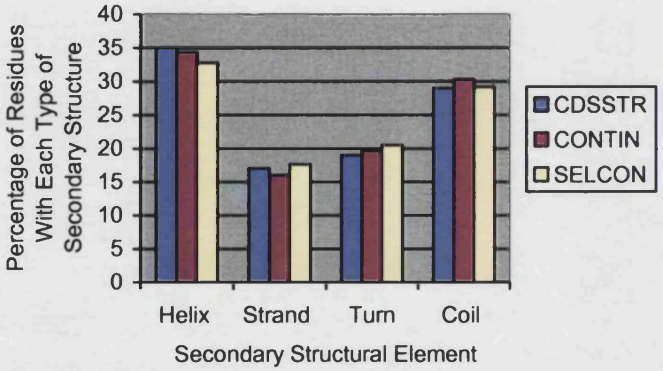



SELCON3

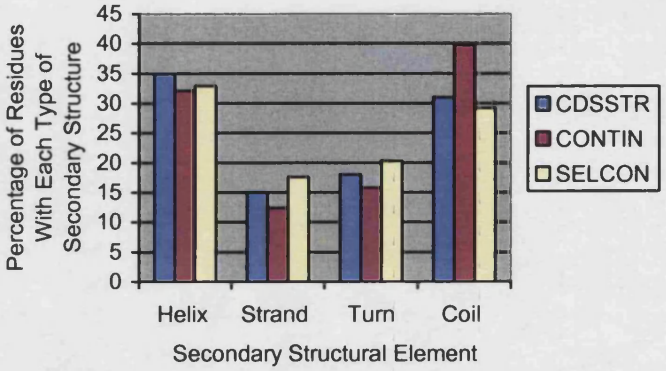



Graphical analyses of the CD spectra recorded for $LH_N/A(H^{227}Y)$ using CDSSTR, CONTINLL and SELCON3. The panels on the left represent spectra analysed with protein reference set 3 and those on the right were analysed with reference set 6..

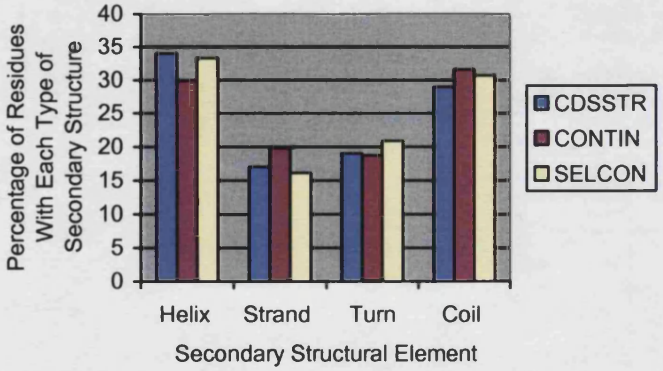
Secondary Structural Composition of LHN/A(H227Y) at pH 7.0 Analysed Using Protein Reference Set 3



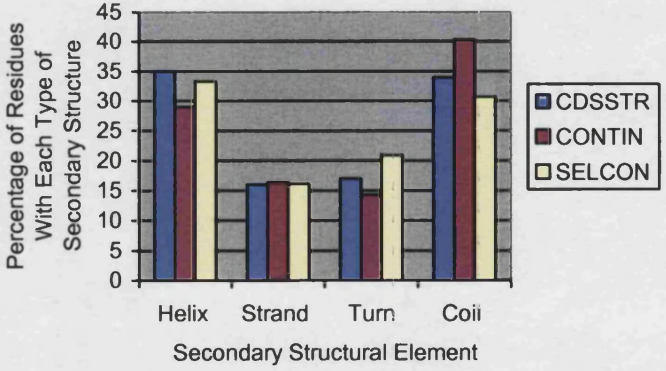

Secondary Structural Composition of LHN/A(H227Y) at pH 7.0 Analysed Using Protein Reference Set 6




Secondary Structural Composition of LHN/A(H227Y) at pH 5.0 Analysed Using Protein Reference Set 3




Secondary Structural Composition of LHN/A(H227Y) at pH 5.0 Analysed Using Protein Reference Set 6



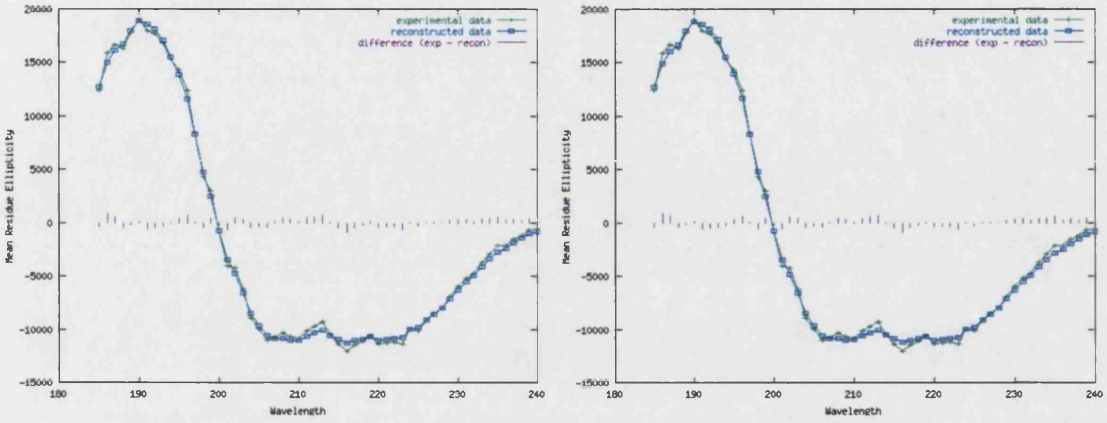
**Secondary Structure of $LH_N/A(H^{227}Y)$ at pH 7.0 and pH 5.0**

| Analysis Program | Reference Set | pH | α-Helix (%) | β-Strand (%) | β-Turn (%) | Random Coil (%) |
|---|---|---|---|---|---|---|
| CDSSTR | 3 | 7 | 35 | 17 | 19 | 29 |
| CONTINLL | 3 | 7 | 34.4 | 16 | 19.7 | 30.3 |
| SELCON3 | 3 | 7 | 32.8 | 17.6 | 20.5 | 29.2 |
| CDSSTR | 6 | 7 | 35 | 15 | 18 | 31 |
| CONTINLL | 6 | 7 | 32.1 | 12.4 | 15.8 | 39.8 |
| SELCON3 | 6 | 7 | 32.9 | 17.6 | 20.3 | 29.2 |
| CDSSTR | 3 | 5 | 34 | 17 | 19 | 29 |
| CONTINLL | 3 | 5 | 29.9 | 19.8 | 18.7 | 31.6 |
| SELCON3 | 3 | 5 | 33.3 | 16.1 | 20.9 | 30.7 |
| CDSSTR | 6 | 5 | 35 | 16 | 17 | 34 |
| CONTINLL | 6 | 5 | 29 | 16.4 | 14.3 | 40.3 |
| SELCON3 | 6 | 5 | 33.3 | 16.1 | 20.9 | 30.7 |

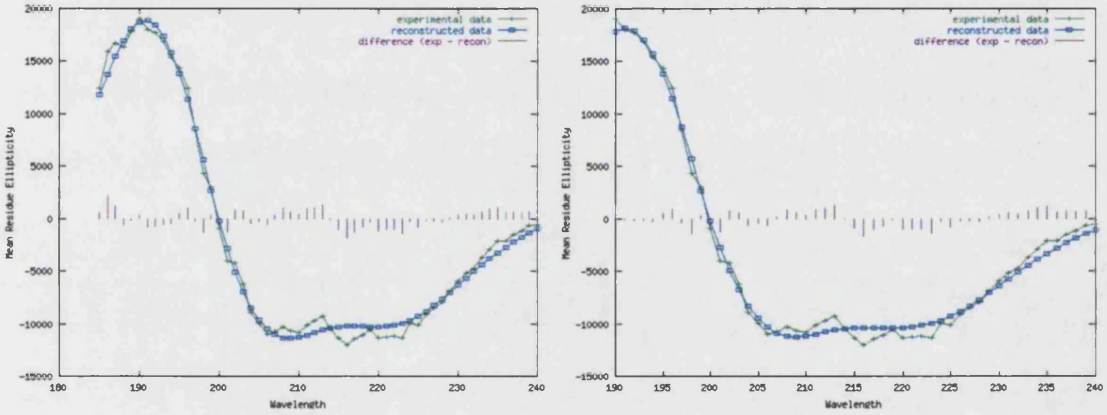
The percentage of residues involved in secondary structural elements (helix, strand, turn and random coil), as analysed using the programs CDSSTR, CONTINLL and SELCON3 with protein reference sets 3 and 6, are presented.

**Analysis of Far-UV CD Spectra Recorded for $LH_N/A(H^{227}Y)$ at pH 5.0**

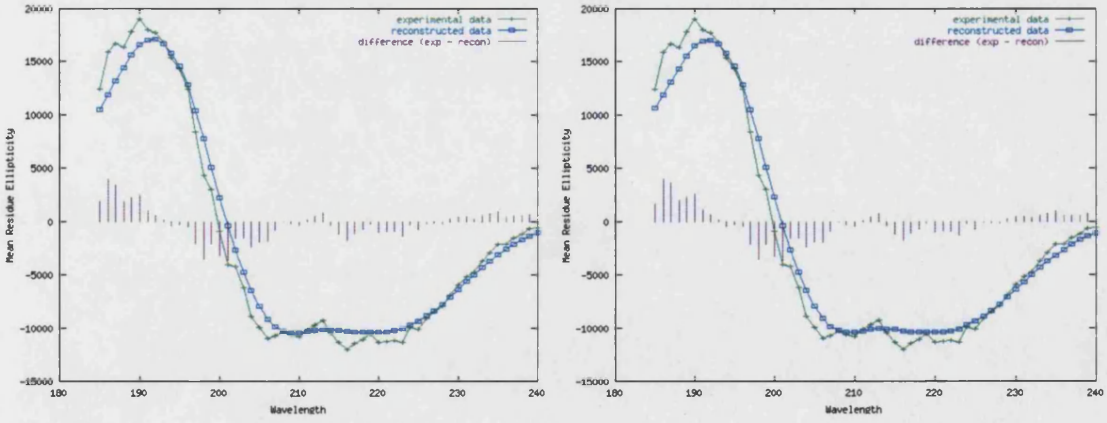
CDSSTR




CONTINLL



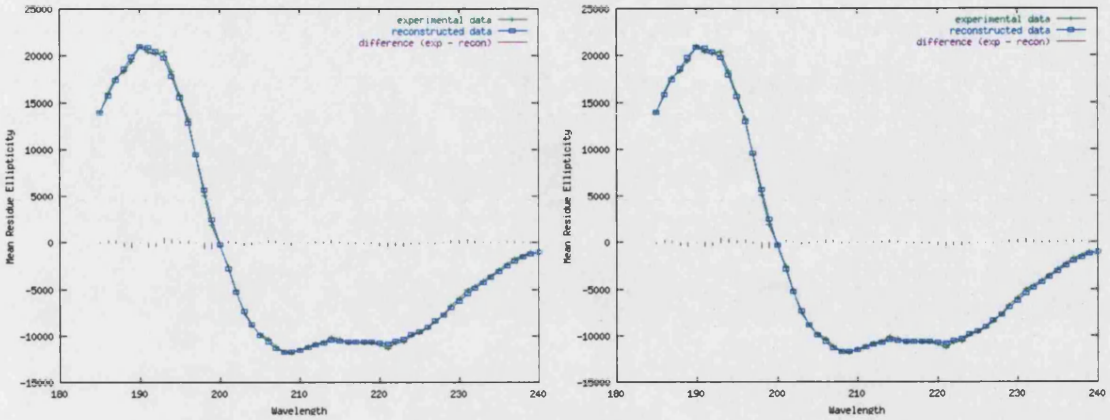

SELCON3



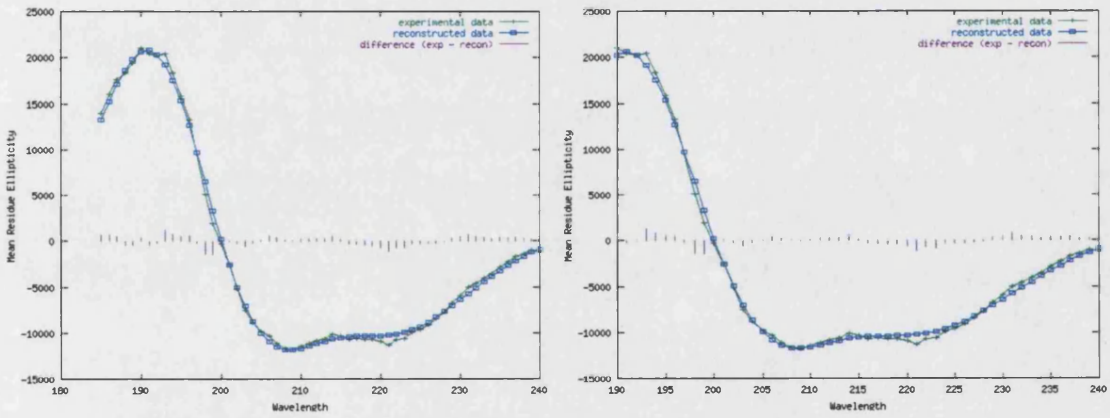

Graphical analyses of the CD spectra recorded for $LH_N/A(H^{227}Y)$ at pH 5.0 using CDSSTR, CONTINLL and SELCON3. Panels on the left represent those analysed with protein reference set 3 and those on the right with reference set 6.

**Analysis of Far-UV CD Spectra Recorded for $LH_N/A(H^{227}Y)$ at pH 7.0**

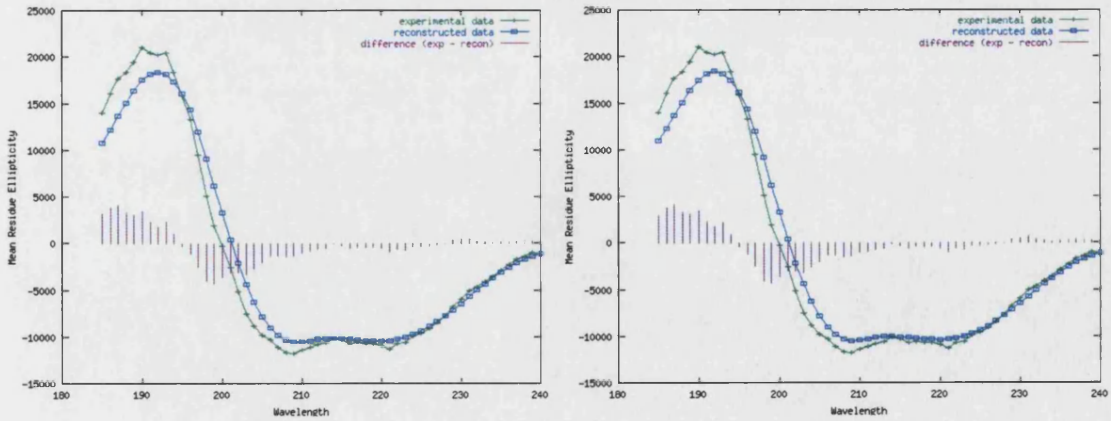
CDSSTR




CONTINLL




SELCON3




Graphical analyses of the CD spectra recorded for $LH_N/A(H^{227}Y)$ at pH 7.0 using CDSSTR, CONTINLL and SELCON3. The panels on the left represent those spectra analysed with protein reference set 3 and those on the right with reference set 6.

# APPENDIX FIVE
# BRIEF DESCRIPTION OF COMPUTER PROGRAMS

**PROGRAMS FOR SEQUENCE ANALYSIS**

**PROGRAMS FOR PROCESSING X-RAY DIFFRACTION DATA**

**PROGRAMS FOR MOLECULAR REPLACEMENT**

**PROGRAMS FOR MODEL BUILDING AND REFINEMENT**

**IMAGING PROGRAMS**

**PROGRAMS FOR STRUCTURAL ANALYSIS**

**PROGRAMS FOR CD SPECTRUM ANALYSIS**

## PROGRAMS FOR SEQUENCE ANALYSIS:

**CLUSTAL W (version 1.8)** (Thompson et al. 1994)

www.ebi.ac.uk/clustalw/

This is a web-based, general purpose multiple alignment program for proteins or DNA. Clustal W can perform multiple sequence alignments and profile alignments and can also be used to produce phylogenetic trees.

**PROTPARAM** (Gill and von Hippel 1989)

http://ca.expasy.org/tools/#protparam.html

A web-based tool for the computation of various physical and chemical parameters (including molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index and grand average of hydropathicity) for a given protein sequence.

## PROGRAMS FOR PROCESSING X-RAY DIFFRACTION DATA:

**HKL2000** (Otwinowski and Minor 1997)

The HKL Suite is a package of programs intended for the analysis of X-ray diffraction data collected from single crystals. It consists of three programs: XDisplayF for visualisation of the diffraction pattern; Denzo for data reduction and integration; and Scalepack for merging and scaling of the intensities.

**DENZO and XDISPLAYF** (Otwinowski and Minor 1997)

These are programs for the visualisation and preliminary analysis of original, unprocessed diffraction data. Diffraction patterns are indexed, crystal and detector parameters are refined and diffraction maxima are integrated. Denzo takes raw X-ray diffraction data and reduces them to a file containing the h,k,l indexes and background and L-P corrected intensity of the spots on the image, and an estimate of the error.

**SCALEPACK** (Otwinowski and Minor 1997)

Scalepack finds the relative scale factors between measurements, precisely refines crystal parameters using the entire data set and performs merging and statistical analysis of measurements related by space group symmetry.

**SCALEPACK2MTZ and TRUNCATE**

This program (part of the CCP4 Suite) converts the merged data output from Scalepack into MTZ format. After the conversion, the Truncate program (also part of the CCP4 program suite) is used to convert the intensities to structure factor amplitudes. Truncate obtains structure factor amplitudes and/or generates useful intensity statistics. Truncate is normally used to read a file of averaged intensities and write a file containing mean amplitudes and the original intensities. The amplitudes are put on an approximate absolute scale using the scale factor taken from a Wilson plot.

**MTZ2XPLOR**

The Mtz2various program is part of the CCP4 Suite and is used to convert the output from Scalepack2mtz into a variety of other formats. Mtz2xplor reads the MTZ file and produces an ASCII file in a suitable form for X-PLOR/CNS. The ASCII file may contain amplitudes, intensities or differences.

## PROGRAMS FOR MOLECULAR REPLACEMENT

**6D_MOLEMAN**

A program for manipulation and analysis of PDB files. Applications include chain renumbering (two digits) and truncation of amino acid side chains to alanine and glycine.

**MATTHEWS_COEF**

Part of the CCP4 Suite, this program calculates the Matthews coefficient (Vm) and solvent content of the crystal from the unit cell dimensions (in Ångstroms) and the molecular weight (in Daltons) of the molecules in the unit cell.

**AMORE** (Navaza 1994)

AMoRe contains a set of programs for molecular replacement: ROTING determines the rotations that superimpose a search molecule upon the homologous ones within the target crystal by calculating the overlap within a conveniently chosen region of volume of the observed Patterson function and a rotated version of the Patterson function corresponding to the isolated search molecule. TRAING selects the possible translations of an oriented model. FITING performs rigid-body refinement.

Other programs reformat the observed data from the new crystal form, and generate and tabulate structure factors from the model in a large P1 cell.

## PROGRAMS FOR MODEL BUILDING AND REFINEMENT

**MAKE_CV** (Brunger 1992)

Part of the CNS Suite of programs, make_cv sets up a test array for cross-validation ($R_{free}$) using a random selection of data (usually 5-10% of the recorded data).

**GENERATE and GENERATE_EASY**

These programs are both part of the CNS Suite. Generate_easy generates coordinate and structure files for simple, commonly encountered protein models (disulphide bonds are automatically determined by distance). Generate can be used to generate a structure file for a model containing protein, DNA/RNA, water, ligands and/or carbohydrate (i.e. more complex than models that can be handled by generate_easy).

**RIGID** (Brunger 1992; Pannu and Read 1996; Adams et al. 1997)

Part of the CNS Suite of programs, rigid performs crystallographic rigid body refinement, which is usually the first step in the refinement procedure. A rigid geometry is assigned to parts of the structure (or to the whole molecule) and the parameters of these regions are refined as a rigid group (instead of refining the individual atomic parameters).

**MINIMIZE** (Brunger 1992; Pannu and Read 1996; Adams et al. 1997)

Part of the CNS Suite, the minimize program performs crystallographic conjugate gradient minimization refinement to reduce differences between calculated and observed reflection amplitudes.

**BINDIVIDUAL** (Brunger 1992; Pannu and Read 1996)

Part of the CNS Suite of programs, bindividual performs restrained, individual B-factor refinement to assign a temperature factor to each atom in the model.

**ANNEAL** (Brunger et al. 1987; Brunger et al. 1990; Brunger 1992; Rice and Brunger 1994; Pannu and Read; Adams et al. 1997)

Part of the CNS Suite, this program performs crystallographic simulated annealing refinement. Anneal computes a cross-validated estimate of σ-A, determines the weighting scheme between the X-ray refinement target function and the geometric energy function, refines a flat bulk solvent model and an overall anisotropic B value of the model by least-squares minimisation and subsequently refines the atomic positions by simulated annealing. The refinement can be run as either slow-cooling or constant-temperature simulated annealing refinement.

**MODEL-MAP** (Read 1986; Kleywegt and Brunger 1996; Brunger et al. 1997)

Part of the CNS Suite, the model_map programs prepares electron density maps using phase information from a model. The program includes a choice of coefficients and the option to average the structure factors from several models.

**CIS_PEPTIDE**

Part of the CNS Suite of programs this program can be used to identify cis-peptides in a protein structure. The program writes out an appropriate parameter file to maintain cis-peptide geometry during refinement.

**WATER_PICK** (Read 1986; Kleywegt and Brunger 1996)

Part of the CNS Suite, the water_pick program picks water molecules in an electron density map.

**MAPMAN** (Kleywegt and Jones 1996)

This is a program for manipulation and analysis of electron density maps. Applications within Mapman include conversion of maps produced by different programs; producing DSN6 map files for O; and simple mathematics on maps.

**O (version 8.0)** (Jones et al. 1991)

O is a general purpose macromolecular modelling environment in which the electron density maps can be displayed and the structural model viewed and edited. Amino acid residues can repositioned, mutated or rebuilt to fit the observed electron density, bonds can be drawn and distances between pairs of atoms determined. The model can also be displayed in several ways (e.g. ball-and stick).

## IMAGING PROGRAMS

**MOLSCRIPT** (Kraulis 1991)

A program for creating schematic or detailed molecular graphics images from molecular 3D coordinates. The program reads a PDB file and displays the structure in terms of its secondary structural elements, which can then be edited and specifically coloured to produce a molecular image. Specific atoms and/or amino acid side chains can be selected and displayed.

**POVRAY**

http://www.povray.org/

The "Persistence of Vision Raytracer" is a high-quality tool for creating three-dimensional graphics. It can display a Molscript file and allow manipulation of the data to produce a high-quality image. The images can be sized and output in a number of formats.

**GIMP**

http://www.gimp.org/

The "GNU Image Manipulation Program" can be used for image authoring and composition. GIMP can be used to edit and label images rendered using Povray. The edited images can then be output in a number of formats.

**GRASP** (Nicholls et al. 1991)

"Graphical Representation and Analysis of Surface Properties"

GRASP is a graphics program written for Silicon Graphics computers that is used to visualise macromolecules. Its strength compared to other such programs is its facility with surfaces and electrostatics.

## PROGRAMS FOR STRUCTURAL ANALYSIS

**PROCHECK** (Laskowski et al. 1993)

Part of the CCP4 Suite, Procheck contains a set of programs that check the stereochemical quality of a protein structure and produce a number of Postscript plots analysing its overall and residue-by-residue geometry. CLEAN produces a "cleaned-up" version of the coordinates file. SECSTR calculates all the required torsion angles,

main-chain hydrogen-bond energies and secondary structure assignments. The NB program identifies non-bonded interactions between different pairs of residues in the protein structure. ANGLEN calculates all main-chain bond lengths and bond angles in the structure and the root mean squared distances from a best-fit plane for all planar side chain groups. Finally, the TPLOT, PPLOT and BPLOT programs create the plots and residue-by-residue listing for the structure. The plots created include a Ramachandran plot; chi1 and chi2 plot; rmsd from planarity; and distorted geometry.

**CONTACT and H-BONDS**

This program (part of the CCP4 Suite) computes various types of contacts in protein structures and can also analyse water hydrogen bonding.

**MODEL_STATS**

Part of the CNS Suite of programs, model_stats produces a list of statistics about a crystallographic model. This includes R factors, B factors, deviation from ideal bond lengths and bond angles, solvent content and Luzzati coordinate error.

**LSQKAB (superpose)**

Part of the CCP4 Suite, Lsqkab is used to work out and apply various transformations to coordinate files. The program optimises the fit of a subset of atomic coordinates from one file to the same subset of another file and assumes both sets of coordinates are in the PDB format. Lsqkab can be used to calculate the root mean square coordinate difference between two structures.

## PROGRAMS FOR CD SPECTRUM ANALYSIS

**CONTINLL** (Provencher and Glockner 1981; Van Stokkum et al. 1990)

CONTIN analyses the conformations of proteins from far-UV CD spectra using ridge regression analysis. The contribution of each reference spectrum to the analysis is kept small unless it contributes to a good agreement between the theoretical best-fit curve and the raw data. The fit depends upon the choice of proteins in the reference set and an estimate of secondary structural composition may be improved by including denatured proteins in the reference set for the random conformation.

**SELCON3** (Sreerama and Woody 1993; Sreerama et al. 1999)

SELCON uses a self-consistent method to deconvolute protein CD spectra. Proteins in the reference set are arranged in order of increasing rmsd from the CD spectrum to be analysed and those that are least like the spectrum of interest are deleted. Structural prediction improves when the protein analysed is included in the basis set because the solution is biased toward the test protein structure. An initial estimate of protein structure is included in the database and the secondary structure of the protein of interest is then determined. The solution replaces the initial estimate and the process is repeated until self-consistency is attained.

**CDSSTR** (Compton and Johnson 1986; Manavalan and Johnson 1987; Sreerama and Woody 2000)

CDSSTR implements the variable selection method by performing all possible calculations using a fixed number of proteins from the reference set. The algorithm recognises proteins possessing characteristics not reflected by the test protein or proteins not reflecting the characteristics of the test protein and removes them from the basis set. This method probably produces the most accurate analysis results and will produce results where other methods fail to analyse proteins.